INSOFE
Inspire...Educate...Transform.

Inspire…Educate…Transform.

# Foundations of Statistics and Probability for Data Science

## Basic Probability Concepts, Probability Distributions

Dr. Venkatesh Sunkad

August 11, 2018

**MATERIAL CONTENT FROM Dr. SRIDHAR PAPPU**


FLASHBACK
I REMEMBER IT LIKE IT WAS YESTERDAY...
CMCOTT. 04/13/12 #138

# MAXIMUM SECURITY

$50.7 BILLION SPENT FOR DEFENCE DEVELOPMENT IN 2016 PLACES INDIA AMONG WORLD'S TOP FIVE DEFENCE SPENDERS

**INDIA IS** ahead of Saudi Arabia and Russia's expenditure

**THE US,** China and the UK remain the top three defence spenders ahead of India's fourth place

**$46.6 bn** **INDIA SPENT** $46.6 billion last year, as per a report released on Monday

**THE REPORT** said that India is set to overtake UK's budget by 2018

**$1.6 trillion** The worldwide outlook shows that global defence spending rose by 1 per cent to $1.6 trillion this year, against 0.6 per cent in 2015.

## DEFENCE EXPENDITURE

US $622 bn | China $191.7 bn | UK $5.8 bn

Saudi $48.68bn

Russia $48.44 bn

Over the next three years, India will re-emerge as a key growth market for defence suppliers — Craig Caffrey, principal analyst for Asia-Pacific at 'IHS Janes'

# 38/35 in math, physics: In Bihar, some students score more than total

Faryal Rumi | TNN | Updated: Jun 9, 2018, 16:33 IST

A- A+

*Students checking their results on mobile phones after BSEB release of Intermediate results on website in Patn...* Read More

## HIGHLIGHTS

- The Bihar School Examination Board was again in the limelight when some class XII students claimed that they scored higher marks than the total.
- Some others complained that they received marks in papers they never appeared for.

PATNA: Two years after the infamous topper scam, the Bihar School Examination

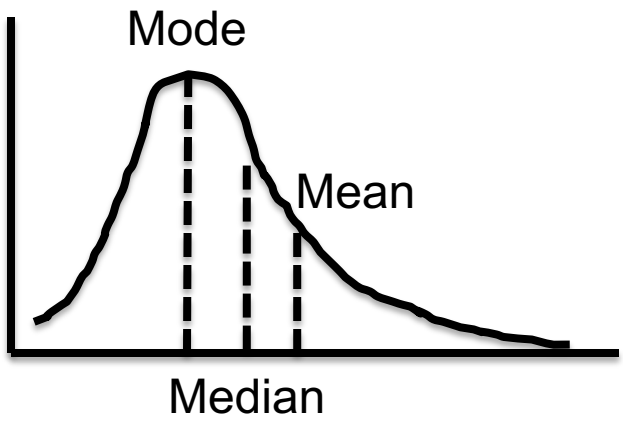
# Data Types – Recent Interview Question

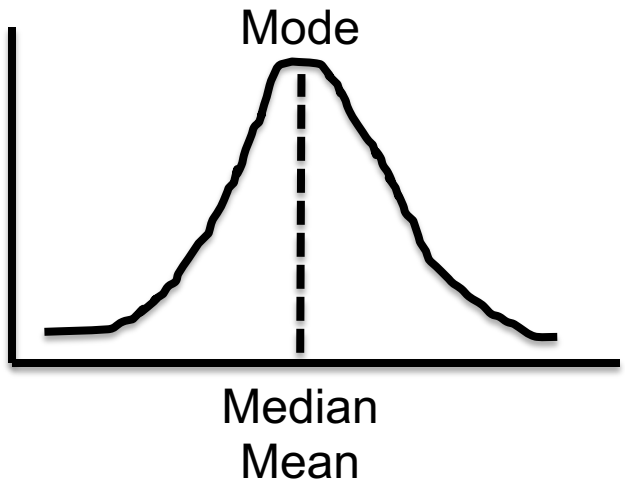
A sample of 400 Bangalore households is selected and several variables are recorded. Which of the following statements is correct?

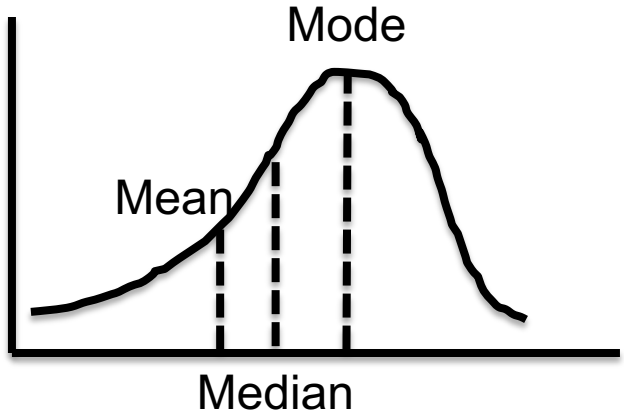
- Socioeconomic status (recorded as “low income”, “middle income”, or “high income”) is nominal level data
- The number of people living in a household is a discrete variable
- The primary language spoken in the household is ordinal level data (recorded as “Kannada”, “Tamil”, etc)

# The Central Tendencies

Identify where the MODE, MEDIAN and MEAN lie in the below distributions.



# Measures of Spread – Recent Interview Question

The spread of the data in a dataset could be studied using _____

- Interquartile range
- Variance
- Standard Deviation
- Range (max-min)
- All of the above

# Measures of Spread – Recent Interview Question

Given the numbers are 68, 83, 58, 84, 100, 64, the second quartile is:

- 74.5
- 75.5
- 75
- 74

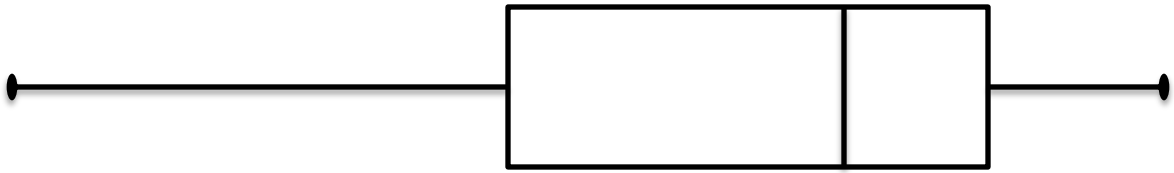
# Measures of Spread – Recent Interview Question

Which of the following plot is used to analyze interquartile range

- Scatterplot
- Histogram
- Lineplot
- Boxplot
- All of the above

# Measures of Spread – Recent Interview Question

What term would best describe the shape of the given boxplot?



- Symmetric
- Skewed with right tail
- Skewed with left tail
- All the above

# Measures of Spread (Dispersion)

Just as Quartiles divide data into 4 equal parts, Deciles divide it into 10 equal parts and Percentiles into 100 equal parts.

Given the above, find the $25^{th}$, $50^{th}$, $75^{th}$ and the $90^{th}$ percentiles for the top 16 global marketing sectors for advertising spending for a recent year according to *Advertising Age*. Also, find Q2, $5^{th}$ decile and IQR. Data in next slide.

| Sector | Ad spending (in $ million) |
|---|---|
| Automotive | 22195 |
| Personal Care | 19526 |
| Entertainment and Media | 9538 |
| Food | 7793 |
| Drugs | 7707 |
| Electronics | 4023 |
| Soft Drinks | 3916 |
| Retail | 3576 |
| Restaurants | 3553 |
| Cleaners | 3571 |
| Computers | 3247 |
| Telephone | 2448 |
| Financial | 2433 |
| Beer, Wine and Liquor | 2050 |
| Candy | 1137 |
| Toys | 699 |

# Measures of Spread (Dispersion)

| Sector | Ad spending (in $ million) |
|---|---|
| Automotive | 22195 |
| Personal Care | 19526 |
| Entertainment and Media | 9538 |
| Food | 7793 |
| Drugs | 7707 |
| Electronics | 4023 |
| Soft Drinks | 3916 |
| Retail | 3576 |
| Cleaners | 3571 |
| Restaurants | 3553 |
| Computers | 3247 |
| Telephone | 2448 |
| Financial | 2433 |
| Beer, Wine and Liquor | 2050 |
| Candy | 1137 |
| Toys | 699 |

$25^{th}$Percentile = 25*(n+1)/100
$25^{th}$Percentile = 25*(16+1)/100
$25^{th}$Percentile = 4.25 ( Between Financial & Telephone)
$25^{th}$Percentile = (2433+2448)/2
**$25^{th}$Percentile (Q1) = 2440.5**

$50^{th}$ Percentile = 50*(n+1)/100
$50^{th}$ Percentile = 50*(16+1)/100
$50^{th}$ Percentile = 8.5 (Cleaners & Retail)
$50^{th}$ Percentile = (3571+3576)/2
**$50^{th}$ Percentile (MEDIAN/Q2) = 3573.5**

**$75^{th}$ Percentile (Q3) = 7750**

**$90^{th}$ Percentile = 20860.5**

**$5^{th}$Decile(Median/Q2)=5*(n+1)/10=3573.5**

**IQR = Q3-Q1 = 7750 – 2440.5 = 5309.5**

# PROBABILITY BASICS


## Sholay

# Probability vs Statistics

- Probability – Predict the likelihood of a future event
- Statistics – Analyse the past events

- Probability – What will happen in a given ideal world?
- Statistics – How ideal is the world?

# Probability vs Statistics


Probability is the basis of inferential statistics.

# Probability - Applications



**Table 1. Life table for the total population: United States, 2003**

| Age | Probability of dying between ages $x$ to $x+1$ | Number surviving to age $x$ | Number dying between ages $x$ to $x+1$ | Person-years lived between ages $x$ to $x+1$ | Total number of person-years lived above age $x$ | Expectation of life at age $x$ |
|---|---|---|---|---|---|---|
| | $q_x$ | $l_x$ | $d_x$ | $L_x$ | $T_x$ | $e_x$ |
| 0–1 | 0.006865 | 100,000 | 687 | 99,394 | 7,743,016 | 77.4 |
| 1–2 | 0.000469 | 99,313 | 47 | 99,290 | 7,643,622 | 77.0 |
| 2–3 | 0.000337 | 99,267 | 33 | 99,250 | 7,544,332 | 76.0 |
| 3–4 | 0.000254 | 99,233 | 25 | 99,221 | 7,445,082 | 75.0 |
| 4–5 | 0.000194 | 99,208 | 19 | 99,199 | 7,345,861 | 74.0 |
| 5–6 | 0.000177 | 99,189 | 18 | 99,180 | 7,246,663 | 73.1 |
| 6–7 | 0.000160 | 99,171 | 16 | 99,163 | 7,147,482 | 72.1 |

Insurance industry uses probabilities in actuarial tables for setting premiums and coverages.

# Probability - Applications

Gaming industry – Establish charges and payoffs

HR – Does a company have biased hiring policies?

Manufacturing/Aerospace – Prevent major breakdowns

# Assigning Probabilities

## Classical Method – *A priori* or Theoretical

Probability can be determined prior to conducting any experiment.

$$P(E) = \frac{\#\ of\ outcomes\ in\ which\ the\ event\ occurs}{total\ possible\ \#\ of\ outcomes}$$

Example: Tossing of a fair die

# Assigning Probabilities

**Empirical Method – *A posteriori* or Frequentist**

Probability can be determined post conducting a thought experiment.

$$P(E) = \frac{\#\ of\ times\ an\ event\ occurred}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occurred}$$

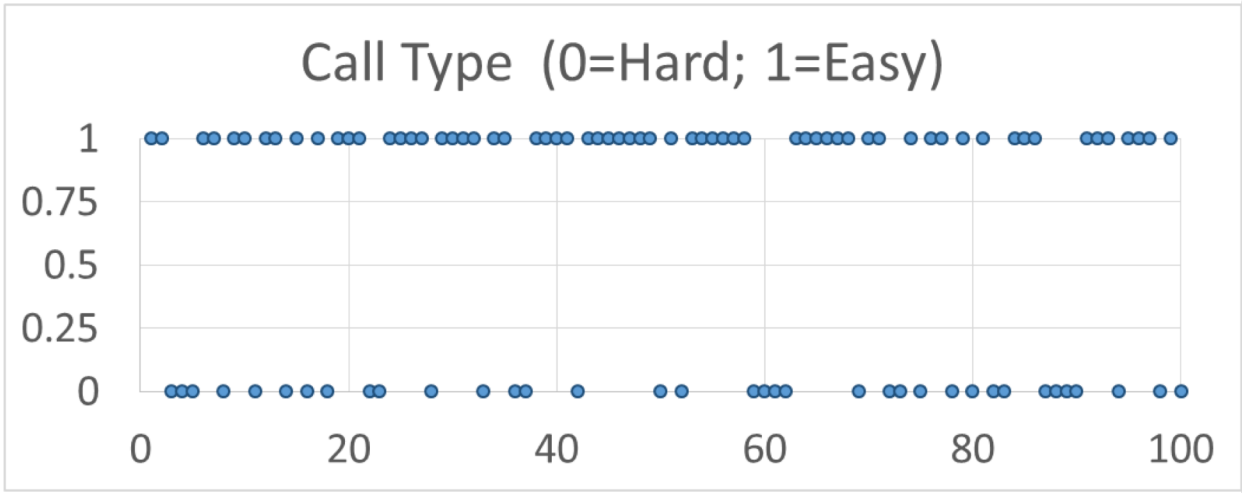
Example: Tossing of a weighted die…well!, even a fair die. The larger the number of experiments, the better the approximation.

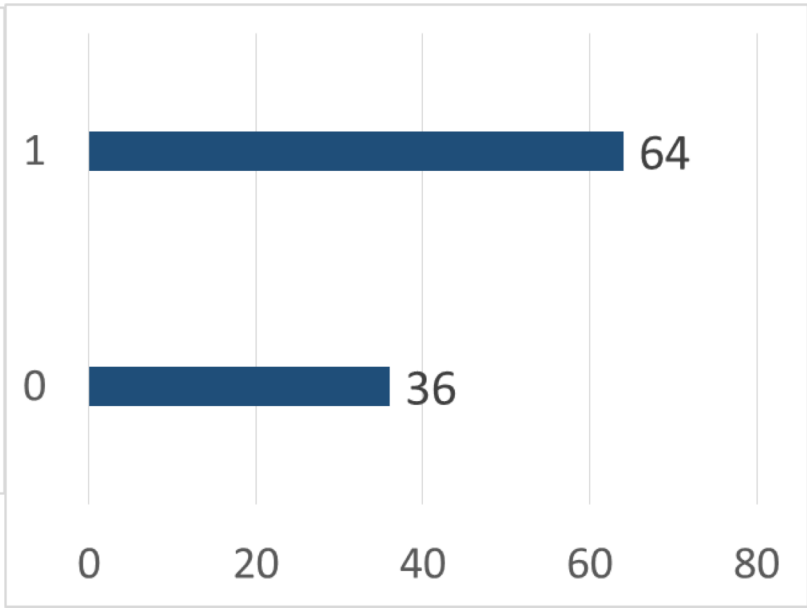
This is the most used method in statistical inference.

# Assigning Probabilities

## Empirical Method – *A posteriori* or Frequentist

100 calls handled by an agent at a call centre

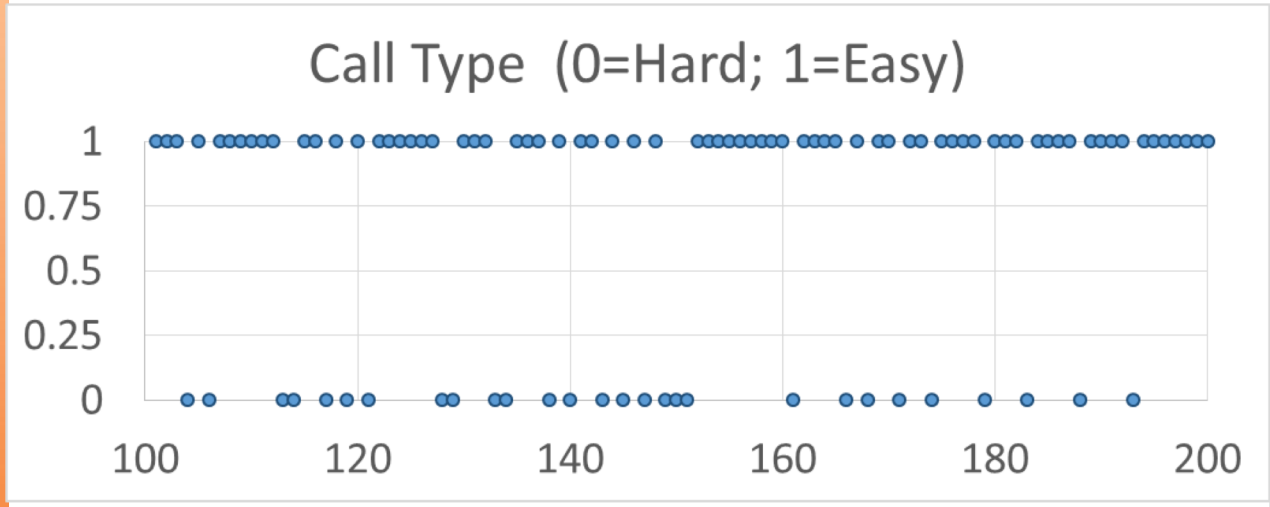


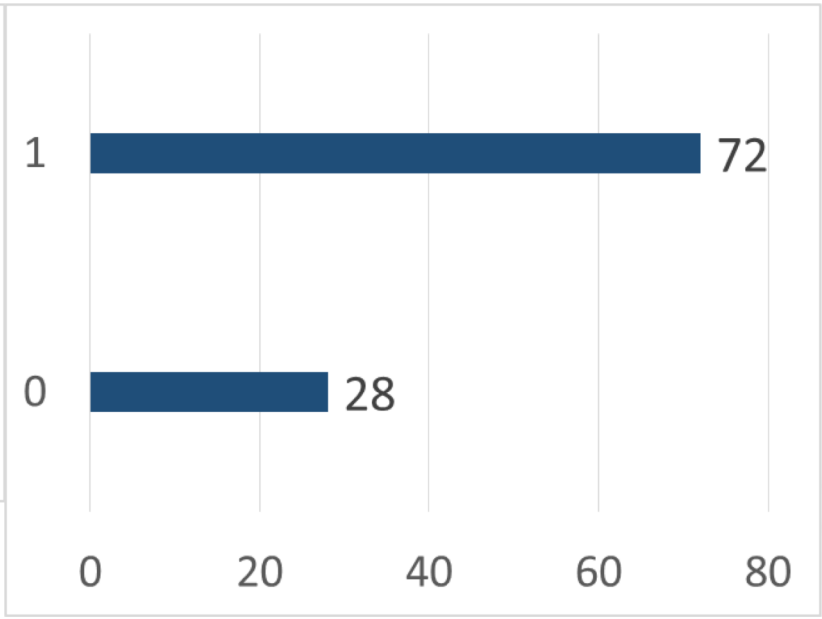



# Assigning Probabilities

## Empirical Method – *A posteriori* or Frequentist

Next 100 calls handled by an agent at a call centre

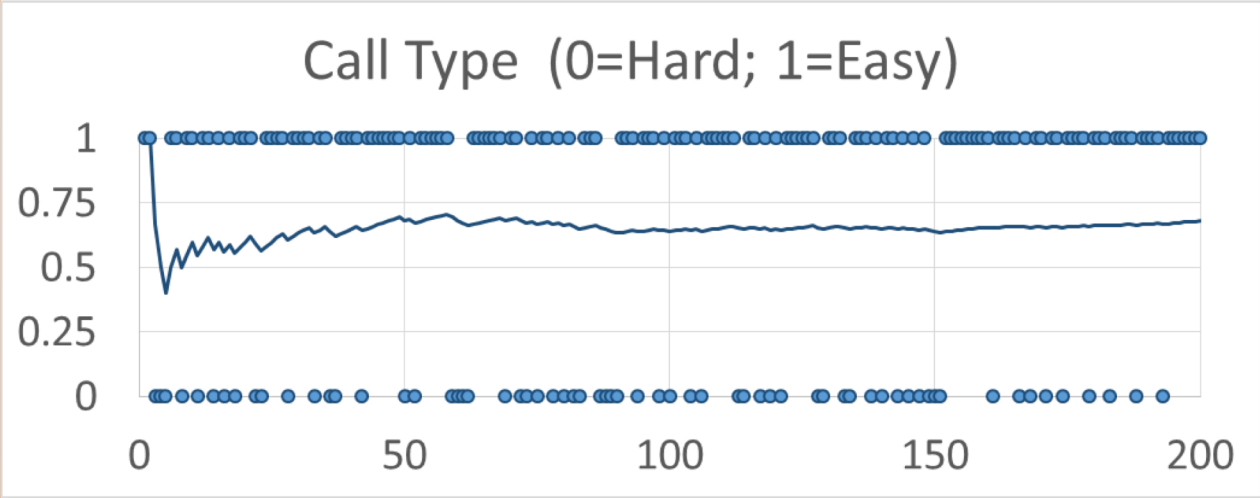


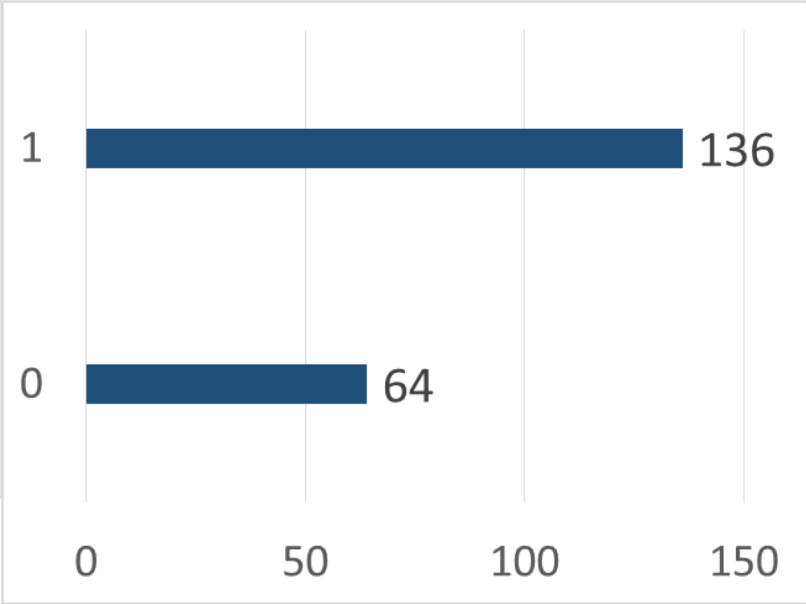



# Assigning Probabilities

## Empirical Method – *A posteriori* or Frequentist

Averages over the long run




P(easy) ~= 0.7

# Assigning Probabilities

## Empirical Method – *A posteriori* or Frequentist

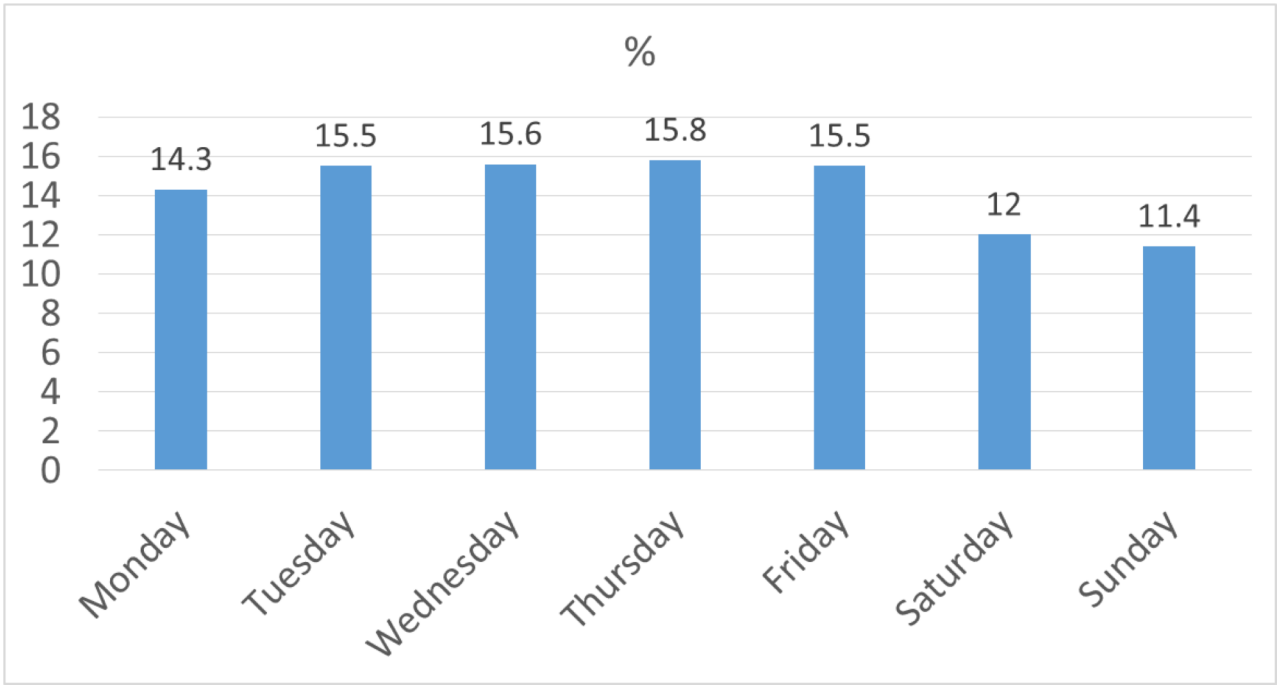
What is the probability of having a monthly income of 1000 BHD?

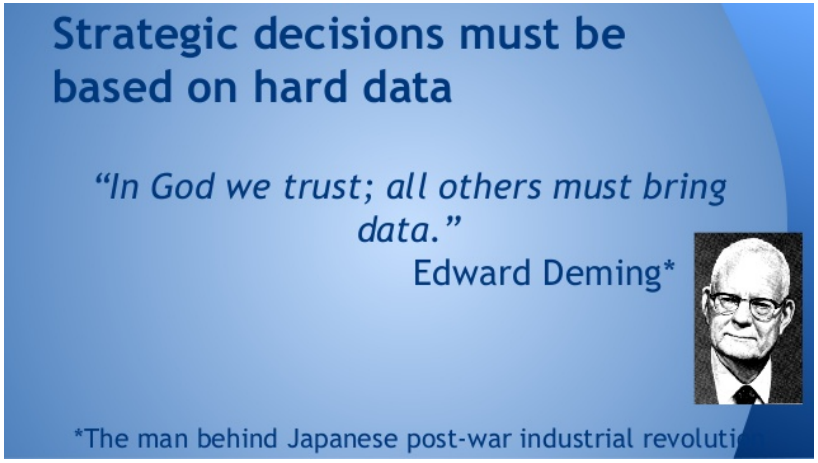
10/23 = 0.43

| INCOME(BHD) | FREQUENCY |
|---|---|
| 100 | 10 |
| 345 | 1 |
| 1000 | 10 |
| 9833 | 2 |

# Assigning Probabilities

What is the probability of a baby being born on a Sunday?



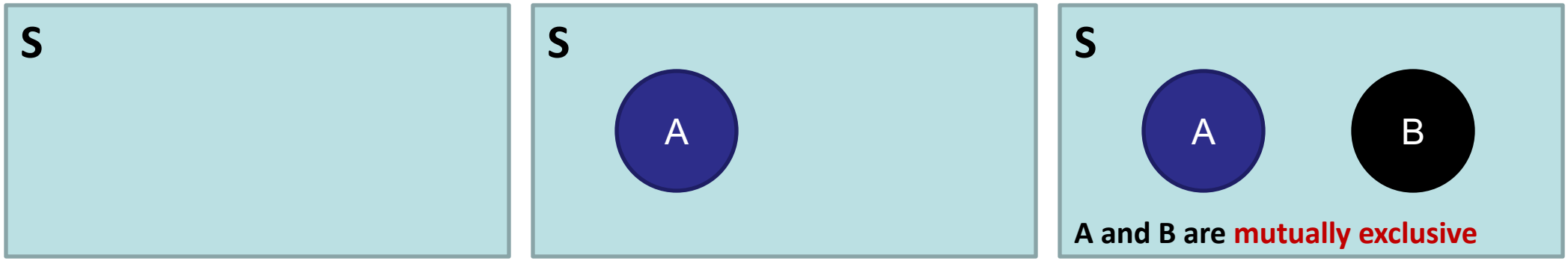

Data from "Risks of Stillbirth and Early Neonatal Death by Day of Week", by Zhong-Cheng Luo, Shiliang Liu, Russell Wilkins, and Michael S. Kramer, for the Fetal and Infant Health Study Group of the Canadian Perinatal Surveillance System. Data of 3,239,972 births in Canada between 1985 and 1998. The reported percentages do not add up to 100% due to rounding.

# Probability - Terminology

Sample Space – Set of all possible outcomes, denoted S.

Event – A subset of the sample space.

# Probability – Rules - Mutually Exclusive




$$P(S) = 1$$

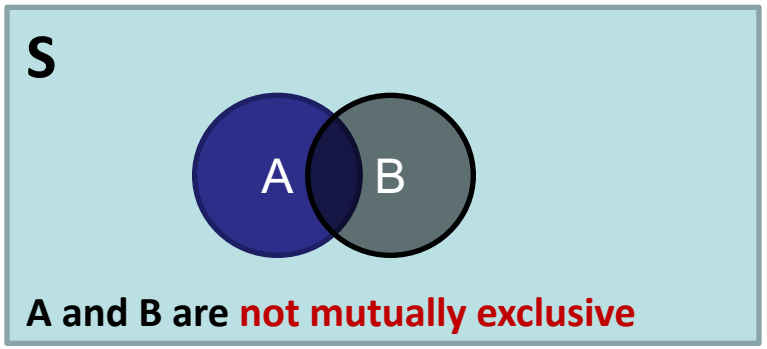
$$0 \leq P(A) \leq 1$$

$$P(A \text{ or } B) = P(A) + P(B)$$

Area of the rectangle denotes sample space, and since probability is associated with area, it cannot be negative.

Mutually Exclusive – If event A happens, event B cannot.

# Probability – Rules




$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

**Example**

Event A – Customers who default on loans

Event B – Customers who are High Net Worth Individuals

# Probability – Rules – Independent Events

Independent Events – Outcome of event B is not dependent on the outcome of event A.

Probability of customer B defaulting on the loan is not dependent on default (or otherwise) by customer A.

$$P(A\ and\ B) = P(A) * P(B)$$

# Probability - Rules

If the probability of getting an *easy* call is 0.7, what is the probability that the next 3 calls will be *easy*?

$$P(easy_1\ and\ easy_2\ and\ easy_3) = 0.7^3 = 0.343$$

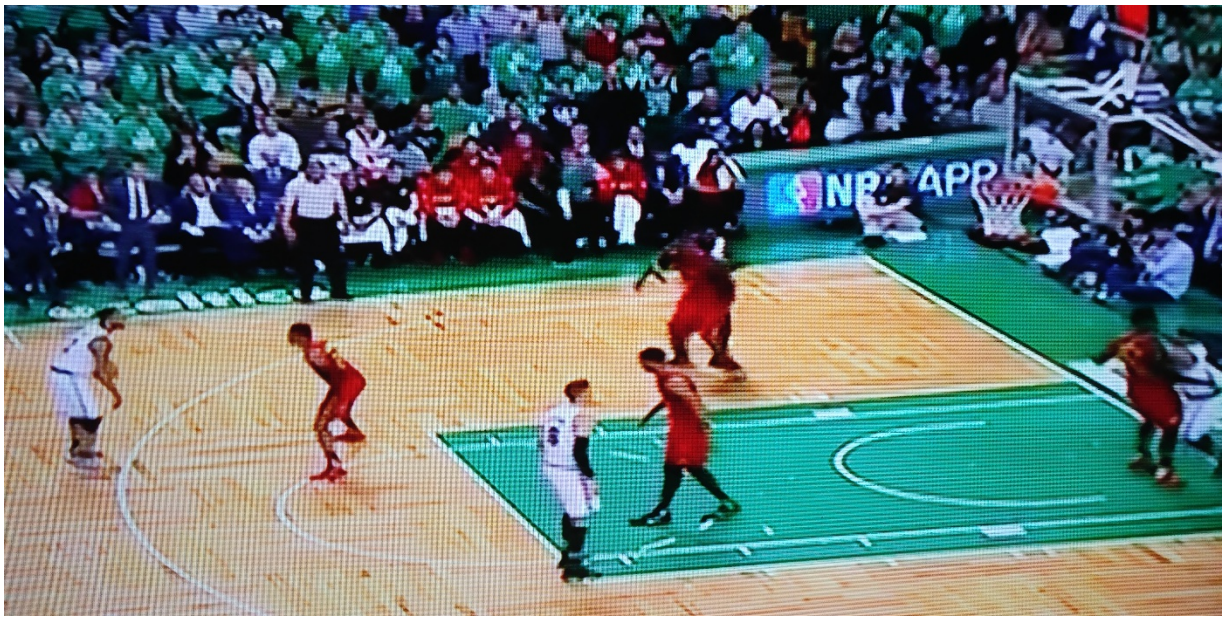
# Probability - Question

A basketball team is down by 2 points with only a few seconds remaining in the game. Given that:

- Chance of making a 2-point shot to tie the game = 50%
- Chance of winning in overtime = 50%
- Chance of making a 3-point shot to win the game = 30%

What should the coach do: go for 2-point or 3-point shot?

What are the assumptions, if any?



# Probability - Question

A basketball team is down by 2 points with only a few seconds remaining in the game. Given that:

- Chance of making a 2-point shot to tie the game = 50%
- Chance of winning in overtime = 50%
- Chance of making a 3-point shot to win the game = 30%

What should the coach do: go for 2-point or 3-point shot?

Ans: **Team goes for 2 point shot then**

P(winning the game) = P(2 Point shot) * P(winning in overtime)

P( winning the game) =1/2 * 1/2 = ¼ =0.25

**Team goes for 3 point shot then**

P(winning the game) = 0.30 **–3 POINT SHOT IS BETTER**

# Probability - Types

Contingency table summarizing 2 variables, *Loan Default* and *Age*:

| | | Age | | | |
|---|---|---|---|---|---|
| | | **Young** | **Middle-aged** | **Old** | **Total** |
| **Loan Default** | **No** | 10,503 | 27,368 | 259 | 38,130 |
| | **Yes** | 3,586 | 4,851 | 120 | 8,557 |
| | **Total** | 14,089 | 32,219 | 379 | 46,687 |

P(Young and Not Defaulting on the loan) = 10503/46687 = 0.225

P(Old and Defaulting on loan) = 120/46687 = 0.003

# Probability - Types

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 10,503 | 27,368 | 259 | 38,130 |
| | Yes | 3,586 | 4,851 | 120 | 8,557 |
| | Total | 14,089 | 32,219 | 379 | 46,687 |

Convert it into probabilities:

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

P(Young and Not Defaulting on the loan) = 10503/46687 = 0.225

P(Old and Defaulting on loan) = 120/46687 = 0.003

P(Yes) = 8557/46687 =0.184

P(Young) = 14089/46687 = 0.302

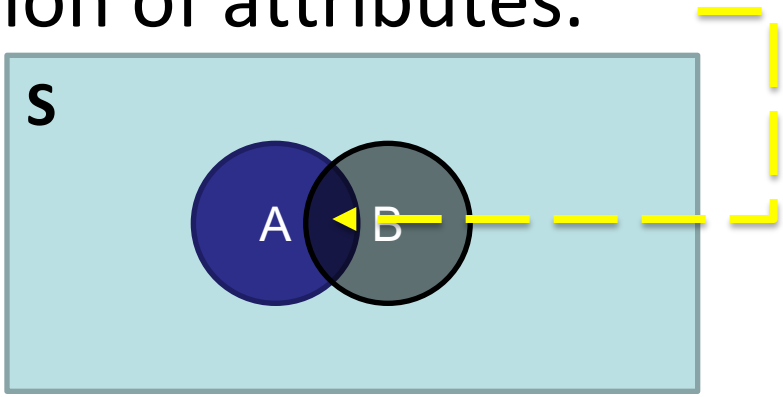
# Probability - Types

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

## Joint Probability

Probability describing a combination of attributes.

P(Yes **and** Young) = 0.077

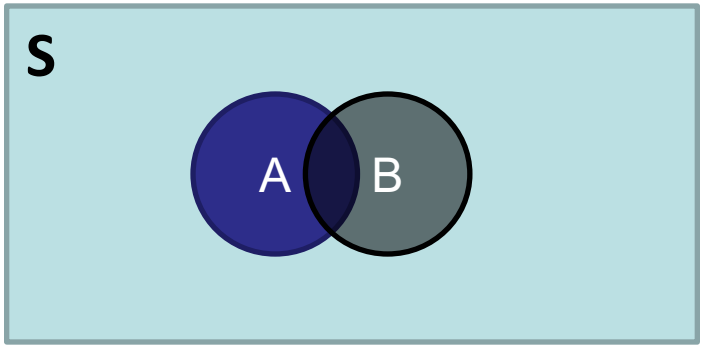


# Probability - Types

## Union Probability

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

P(Yes **or** Young) = P(Yes) + P (Young) – P(Yes and Young) =
0.184 + 0.302 – 0.077 = 0.409



# Probability - Types

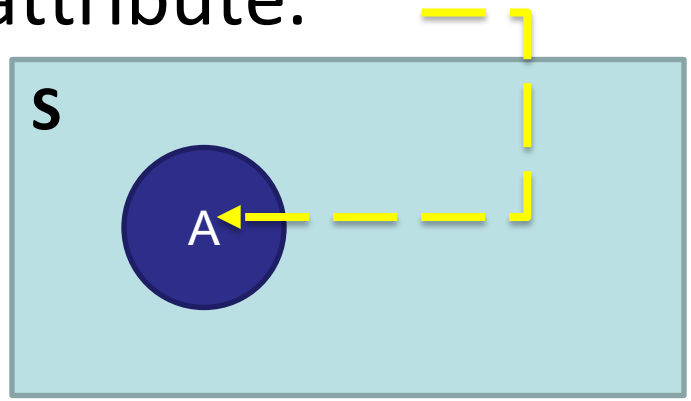
## Marginal Probability

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

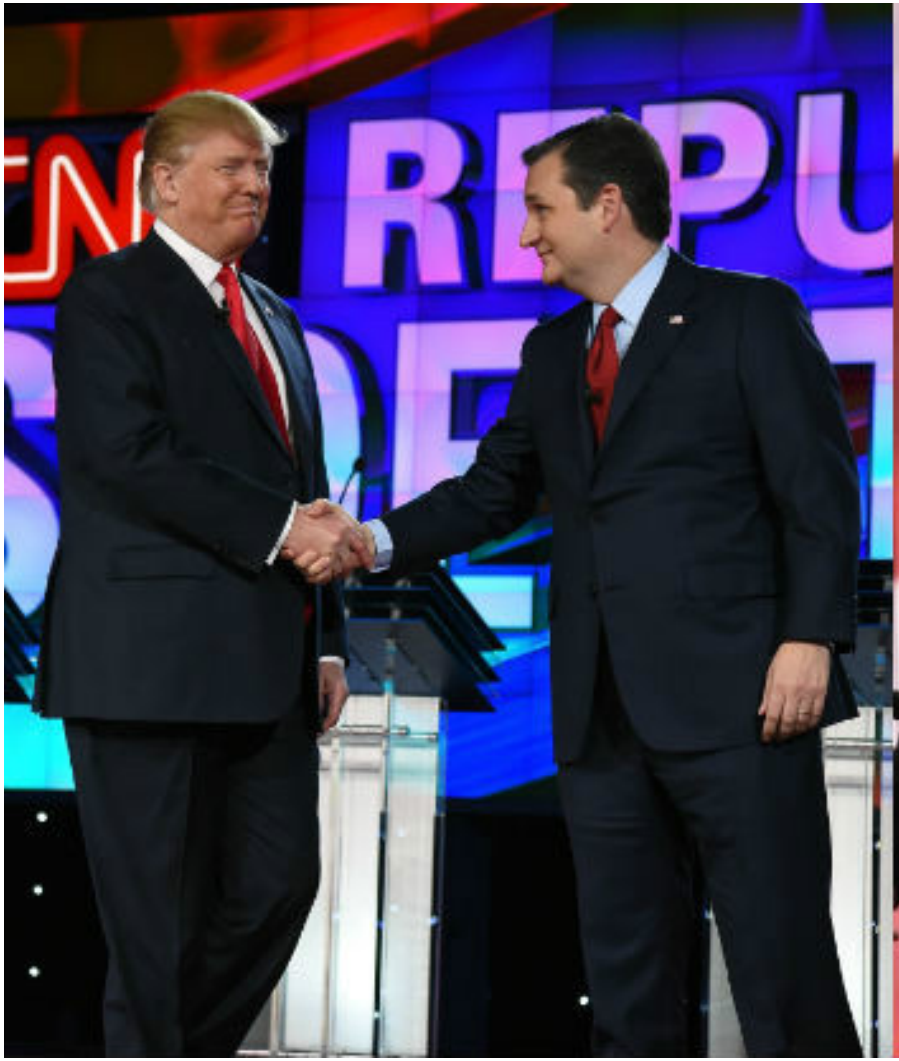
Probability describing a single attribute.

P(No) = 0.816

P(Old) = 0.008



# Independent or Mutually Exclusive?



Donald Trump and Ted Cruz were Republican Party candidates.


Hillary Clinton and Bernie Sanders were Democratic Party candidates.

# Independent or Mutually Exclusive?

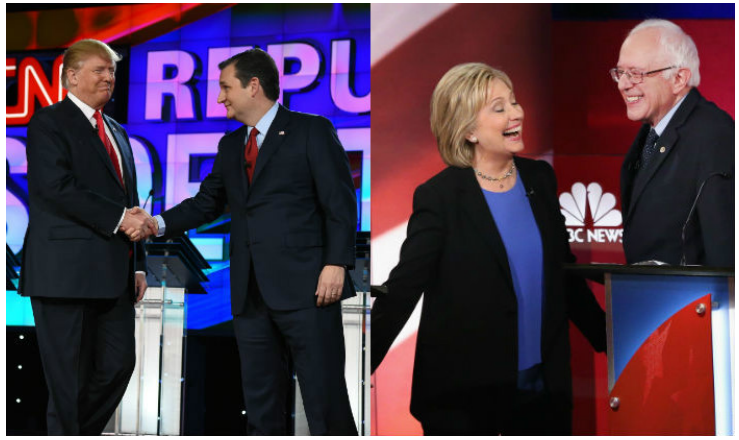
Event A: Trump winning Republican nomination

Event B: Cruz winning Republican nomination

Event C: Clinton winning Democratic nomination

Event D: Sanders winning Democratic nomination

What kinds of events are the below scenarios?



| | |
|---|---|
| Event A and Event B | *Mutually Exclusive* |
| Event C and Event D | *Mutually Exclusive* |
| Event A and Event C | *Independent* |

# Independent or Mutually Exclusive?

Event A: Trump winning Republican nomination

Event B: Cruz winning Republican nomination

Event C: Clinton winning Democratic nomination

Event D: Sanders winning Democratic nomination

Assuming no other candidates are left in the fray and there is a neck-to-neck contest within each party, what is:

P(A and B) *0*

P(B and A) *0*

P(A and C) *½ * ½ = ¼*

P(C and A) *½ * ½ = ¼*

P(A or B) *½ + ½ = 1*

P(B or A) *½ + ½ = 1*

P(A or C) *½ + ½ - ¼ = ¾*

P(C or A) *½ + ½ - ¼ = ¾*

# Probability - Types

BREAK

- Joint Probability
  - P(A and B) = P(A)*P(B)
- Union Probability
  - P (A or B) = P(A) + P(B) – P(A and B)
- Marginal Probability - Probability of a Single Attribute
  - Only one P(A), P(B)
- Conditional Probability

# Probability - Types

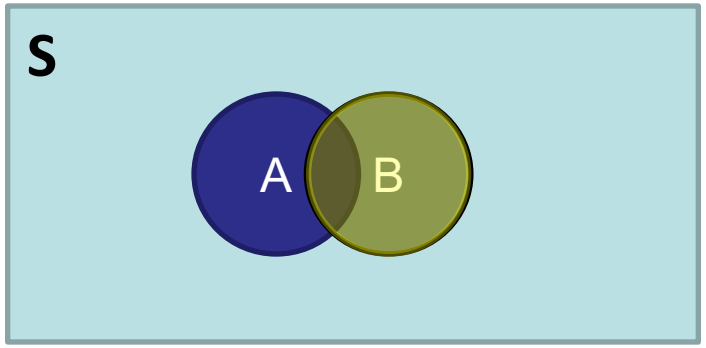
## Conditional Probability

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

Probability of *A* occurring **given that** *B* has occurred.

The sample space is restricted to a single row or column.

This makes rest of the sample space irrelevant.



# Probability - Types

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

## Conditional Probability

What is the probability that a person will not default on the loan payment **given** she is middle-aged?

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 10,503 | 27,368 | 259 | 38,130 |
| | Yes | 3,586 | 4,851 | 120 | 8,557 |
| | Total | 14,089 | 32,219 | 379 | 46,687 |

P(No | Middle-Aged) = ?

# Probability - Types

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 10,503 | 27,368 | 259 | 38,130 |
| | Yes | 3,586 | 4,851 | 120 | 8,557 |
| | Total | 14,089 | 32,219 | 379 | 46,687 |

**Conditional Probability** = $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

Note that this is the ratio of **Joint Probability** to **Marginal Probability**

P(No | Middle-Aged) = $\frac{P(Middle\ aged\ and\ NO)}{P(Middle)} = \frac{0.586}{0.690} = 0.85$

P(No | Middle Aged ) = $\frac{P(Middle\ aged\ and\ NO)}{P(Middle)} = \frac{27368/46687}{32219/46687} = 0.85$

# Conditional Probability – Order Matters

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

**P(No | Middle-Aged) = 0.586/0.690 = 0.85**

What is the probability that a person is middle-aged **given** she has not defaulted on the loan payment?

**P(Middle-Aged | No) = 0.586/0.816 = 0.72 (Order Matters)**

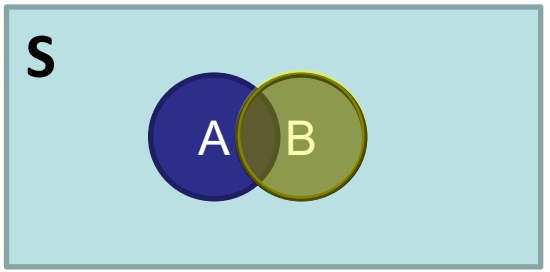
**P(Middle-Aged | No) = 27368/38130 = 0.72 (Order Matters)**

# Probability - Types

## Conditional Probability – Visualizing using Probability Tables and Venn Diagrams

| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 10,503 | 27,368 | 259 | 38,130 |
| | Yes | 3,586 | 4,851 | 120 | 8,557 |
| | Total | 14,089 | 32,219 | 379 | 46,687 |

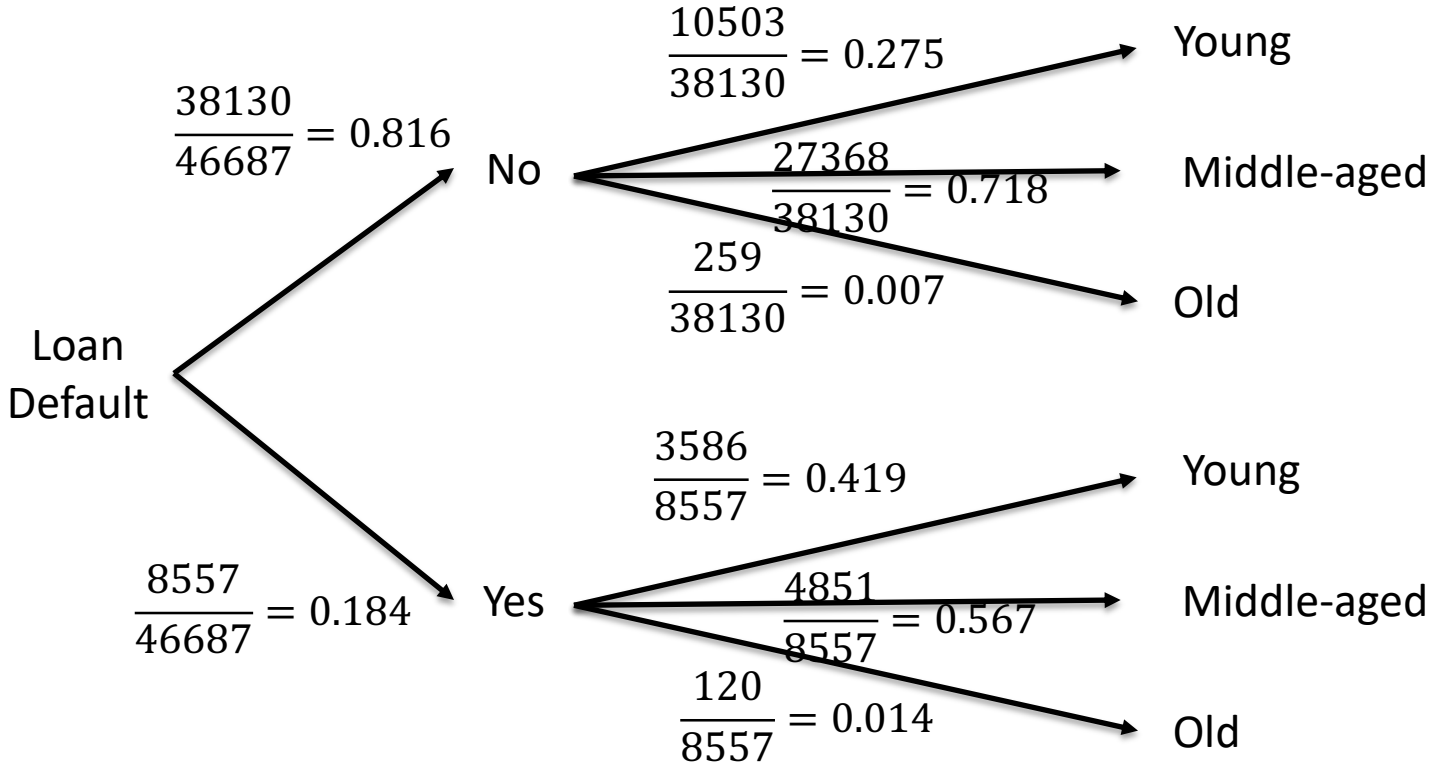
| | | Age | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |



# Probability - Types

## Conditional Probability – Visualizing using Probability Trees

| | | Age (Numbers) | | | | Age (Probabilities) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total | Young | Middle-aged | Old | Total |
| Loan Default | No | 10,503 | 27,368 | 259 | 38,130 | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 3,586 | 4,851 | 120 | 8,557 | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 14,089 | 32,219 | 379 | 46,687 | 0.302 | 0.690 | 0.008 | 1.000 |




Find
P(Young and No)

We know that

$$P(Young/No) = \frac{P(Young\ and\ NO)}{P(No)}$$

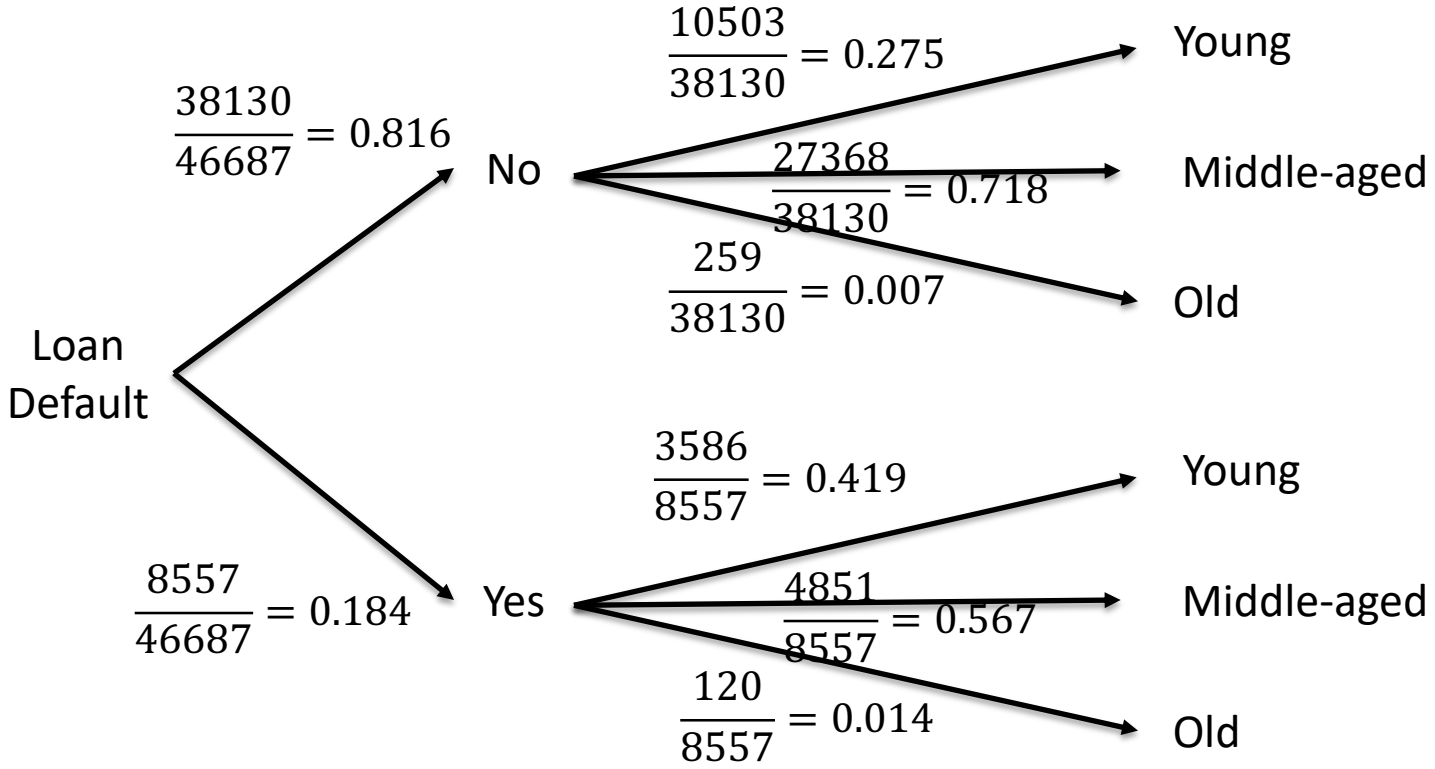
P(Young and No) =
P(Young/No) * P(No)

P(Young and No) =
0.275*0.816 = **0.225**

# Probability - Types

## Conditional Probability – Visualizing using Probability Trees

| | | Age (Numbers) | | | | Age (Probabilities) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total | Young | Middle-aged | Old | Total |
| Loan Default | No | 10,503 | 27,368 | 259 | 38,130 | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 3,586 | 4,851 | 120 | 8,557 | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 14,089 | 32,219 | 379 | 46,687 | 0.302 | 0.690 | 0.008 | 1.000 |




Find

- P(Young and No)
- P(No and Young)
- P(Young) = P(No) P(Young/No) + P(Yes) *P(Young/Yes)
- P(No)
- P(Young | No)
- P(No | Young)

CSE 7315c

# Probability - Types

## Attention Check

Identify the type of probability in each of the below cases:

1. P(Old and Yes)
2. P(Yes and Old)
3. P(Old)
4. P(Yes)
5. P(Old | Yes)
6. P(Yes | Old)
7. P(Young | No)
8. P(Middle-aged or No)
9. P(Old or Young)

| | | Age (Probabilities) | | | |
|---|---|---|---|---|---|
| | | **Young** | **Middle-aged** | **Old** | **Total** |
| **Loan Default** | **No** | 0.225 | 0.586 | 0.005 | **0.816** |
| | **Yes** | 0.077 | 0.104 | 0.003 | **0.184** |
| | **Total** | **0.302** | **0.690** | **0.008** | **1.000** |

1 and 2: **Joint**; 3 and 4: **Marginal**; 5, 6 and 7: **Conditional**; 8 and 9: **Union**

# Probability - Types

## Conditional Probability

$$P(A|B) = \frac{P(A\ and\ B)}{P(B)} \Rightarrow P(A\ and\ B) = P(B) * P(A|B)$$

Similarly

*What happens when A and B are INDEPENDENT?*

$$P(B|A) = \frac{P(A\ and\ B)}{P(A)} \Rightarrow P(A\ and\ B) = P(A) * P(B|A)$$

Equating, we get

$$P(A|B) * P(B) = P(A) * P(B|A)$$

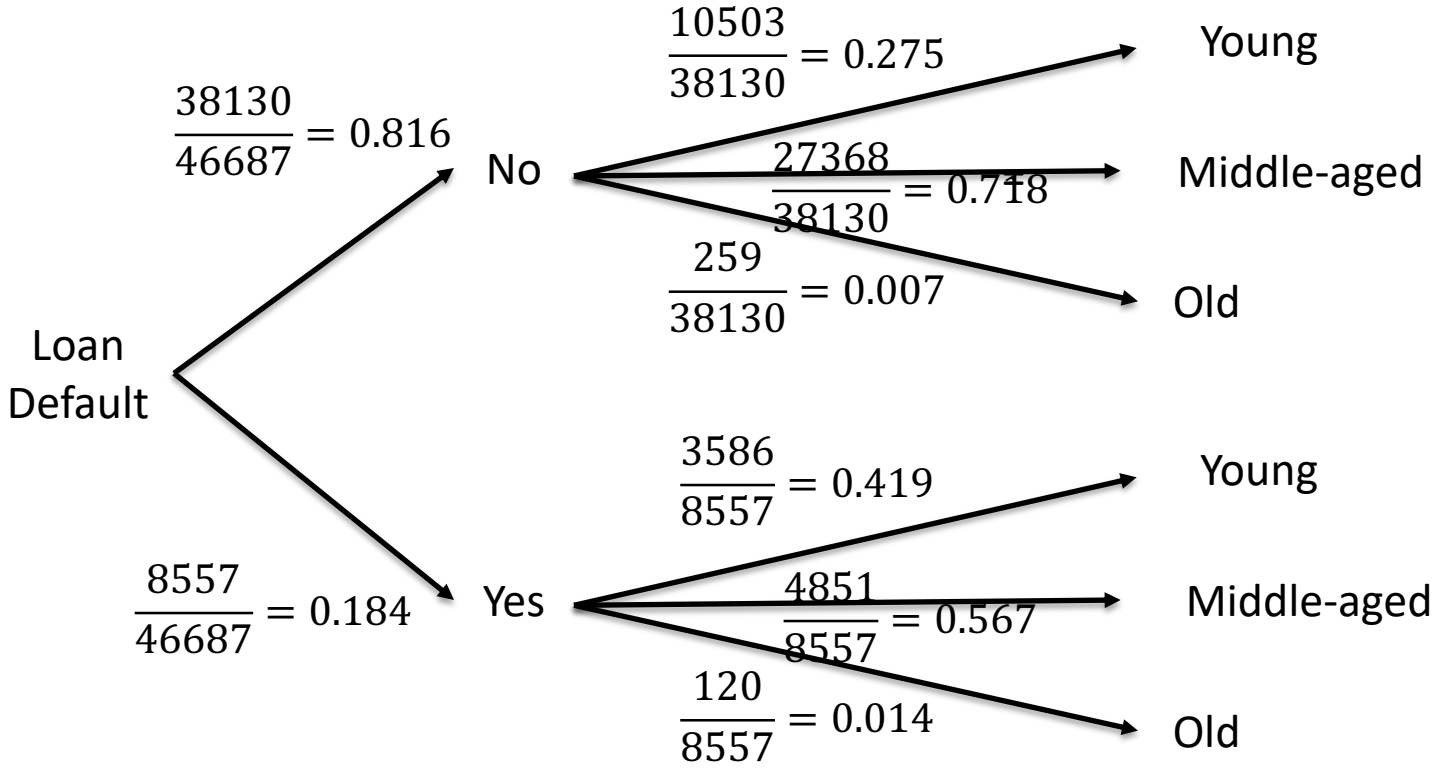
$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

# Probability - Types

## Conditional Probability – Visualizing using Probability Trees

| | | Age (Probabilities) | | | |
|---|---|---|---|---|---|
| | | Young | Middle-aged | Old | Total |
| Loan Default | No | 0.225 | 0.586 | 0.005 | 0.816 |
| | Yes | 0.077 | 0.104 | 0.003 | 0.184 |
| | Total | 0.302 | 0.690 | 0.008 | 1.000 |

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$




**Now find**

P(No | Young)

=

$$\frac{P(No) * P(Young|No)}{P(Young)}$$

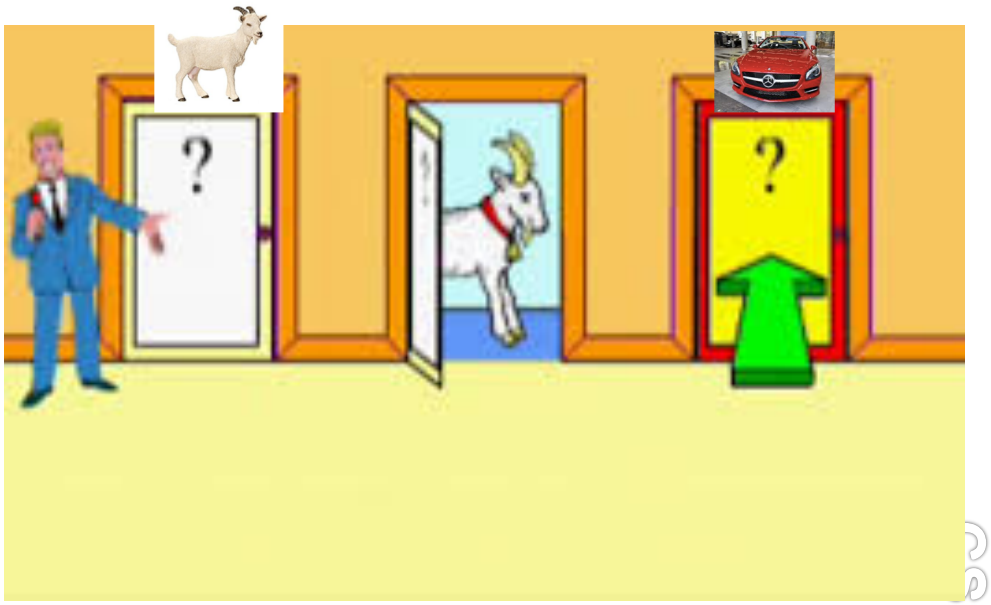
=

$$\frac{0.816 * 0.275}{(0.275*0.816)+(0.419*0.184)}$$

=0.744

# Probability - Types

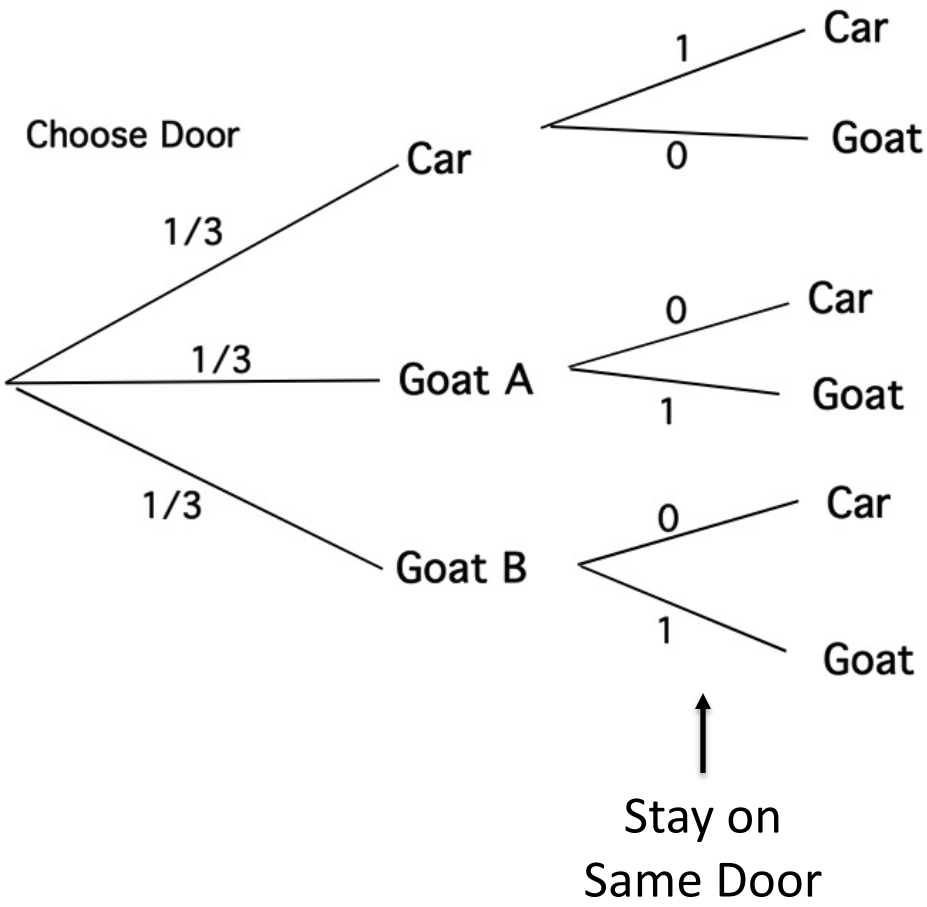
## Monty Hall Problem - Intuitive

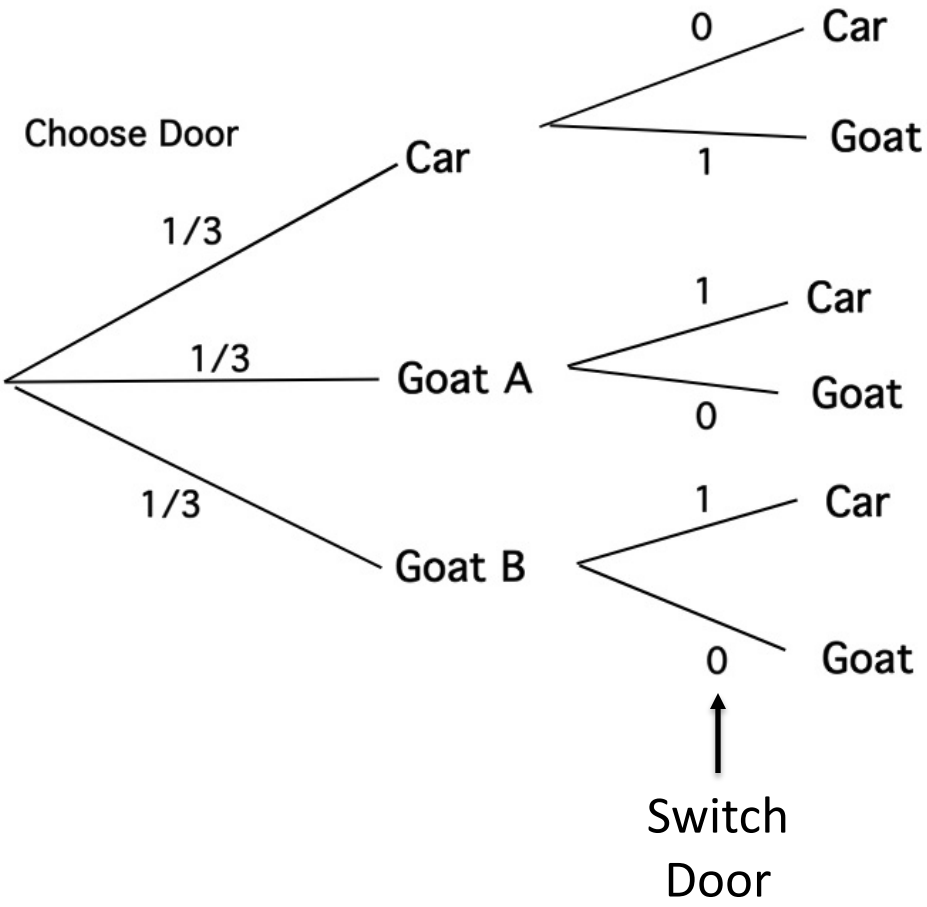




CSE 7315c

# Probability - Types

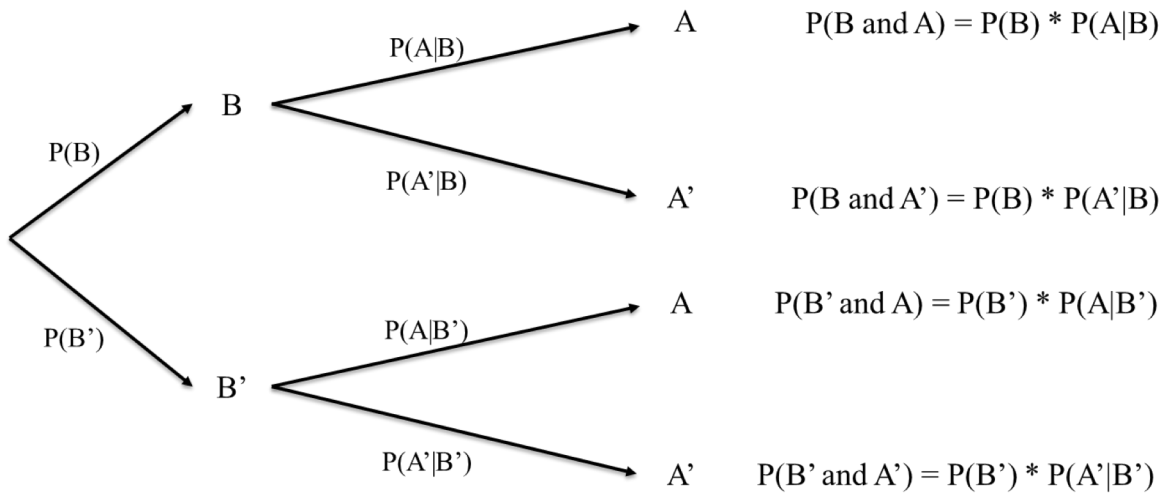
## Monty Hall Problem – Probability Tree



# Probability - Types

## Conditional Probability -> Bayes' Theorem




Note B' means "not B"

$$P(B|A) = \frac{P(B) * P(A|B)}{P(A)} = \frac{P(A|B) * P(B)}{P(A|B) * P(B) + P(A|not\ B) * P(not\ B)}$$

# Bayes' Theorem

Bayes' Theorem allows you to find reverse probabilities, and to allow **revision of original probabilities** with new information.

## Case – Clinical trials

Epidemiologists claim that probability of breast cancer among Caucasian women in their mid-50s is 0.005. An established test identified people who had breast cancer and those that were healthy. A new mammography test in clinical trials has a probability of 0.85 for detecting cancer correctly. In women without breast cancer, it has a chance of 0.925 for a negative result. If a 55-year-old Caucasian woman tests positive for breast cancer, what is the probability that she in fact has breast cancer?

P(Cancer) = 0.005
P(Test positive | Cancer) = 0.85
P(Test negative | No cancer) = 0.925
P(Cancer | Test positive) = ?

# Bayes' Theorem

## Case – Clinical trials

P(Cancer) = 0.005 (*aka* Prior Probability)

P(Test positive | Cancer) = 0.85 (*aka* Likelihood)

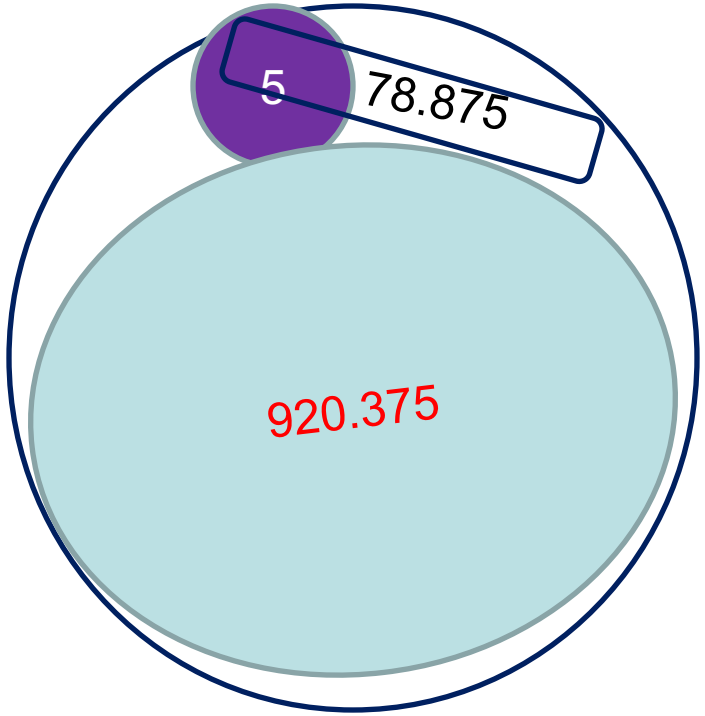
P(Test negative | No cancer) = 0.925

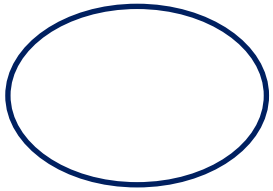
P(Cancer | Test positive) = ? (*aka* Posterior or Revised Probability)

P(Test Positive) *aka* Evidence

$$Posterior\ Probability = \frac{Prior\ Probability * Likelihood}{Evidence}$$

# Cancer Detection – Bayes Theorem

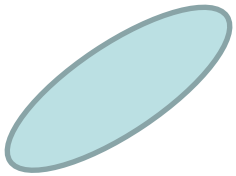


Let us assume 1000 women in the mid 50

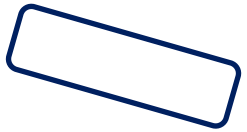

P(Cancer) = 0.005 or 5/1000. So for every 1000 women in their mid 50s, 5 get breast cancer

P(No Cancer) = 1- 0.005 = 0.995 = 995 women



P(test –ve / No cancer) = 0.925 = 92.5% of 995 = 920.375

P( test +ve / No cancer ) = 74.625 (995 – 920.375)



P(Test +ve) = P(test +ve /Cancer) + P(test +ve/No Cancer)
0.85*5 + 74.625 = 78.875

# Bayes' Theorem

## Case – Clinical trials

P(Cancer) = 0.005 (*aka* Prior Probability)

P(Test positive | Cancer) = 0.85 (*aka* Likelihood)

P(Test negative | No cancer) = 0.925

P(Cancer | Test positive) = ? (*aka* Posterior or Revised Probability)

P(Test Positive) *aka* Evidence

$$P(Cancer|Test+) = \frac{P(Cancer) * P(Test+|\ Cancer)}{P(Test+)}$$

$$P(Cancer|Test+) = \frac{P(Cancer) * P(Test+|Cancer)}{P(Test+|Cancer) * P(Cancer) + P(Test+|No\ cancer) * P(No\ cancer)}$$

$$= \frac{0.005 * 0.85}{0.85 * 0.005 + 0.075 * 0.995} = \frac{0.00425}{0.078875} = 0.054$$

**Homework**

Draw a Probability Table and a Probability Tree for the above case.

# Bayes' Theorem

## Case – Spam filtering

Apache SpamAssassin™

**Latest News**

*2015-04-30:* SpamAssassin 3.4.1 has been released! Highlights include:

- improved automation to help combat spammers that are abusing new top level dc
- tweaks to the SPF support to block more spoofed emails;
- increased character set normalization to make rules easier to develop and stop sp
- continued refinement to the native IPv6 support; and
- improved Bayesian classification with better debugging and attachment hashing.

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word "free" appears in 20% of the mails marked as spam, i.e., P(Free | Spam) = 0.20. Assuming 0.1% of non-spam mail includes the word "free" and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word "free" appears in it.

# Bayes' Theorem

BREAK

## Case – Spam filtering

P(Spam) = 0.50

P(Free | Spam) = 0.20

P(Free | No spam) = 0.001

P(Spam | Free) = ?

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free)}$$

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free|Spam) * P(Spam) + P(Free|No\ spam) * P(No\ spam)}$$

$$= \frac{0.5 * 0.2}{0.2 * 0.5 + 0.001 * 0.5} = \frac{0.1}{0.1005} = 0.995$$

This helps the spam filter automatically classify the messages as spam.

A slight detour

# HOW GOOD IS YOUR CLASSIFICATION?

# Confusion Matrix

| Spam filtering | | Predicted | | Total |
|---|---|---|---|---|
| | | Positive | Negative | |
| Actual | Positive | 952 | 526 | 1478 |
| | Negative | 167 | 3025 | 3192 |
| Total | | 1119 | 3551 | 4670 |

| | | Predicted | | METRICS |
|---|---|---|---|---|
| | | Positive | Negative | |
| Actual | Positive | True +ve | False –ve | Recall/Sensitivity/True Positive Rate (Minimize False –ve) |
| | Negative | False +ve | True –ve | Specificity/True Negative Rate (Minimize False +ve) |
| | | Precision | | Accuracy, $F_1$ score |

# Confusion Matrix - Metrics

| | | Predicted | | |
|---|---|---|---|---|
| | | Positive | Negative | |
| **Actual** | Positive | True +ve | False –ve | Recall/Sensitivity/True Positive Rate (Minimize False –ve) |
| | Negative | False +ve | True –ve | Specificity/True Negative Rate (Minimize False +ve) |
| | | Precision | | Accuracy, $F_1$ score |

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{Actual+ve}}$$

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{True+ve + False−ve}}$$

$$\text{Specificity} = \frac{\text{True −ve}}{\text{Actual−ve}}$$

$$\text{Specificity} = \frac{\text{True −ve}}{\text{False+ve +True −ve}}$$

$$\text{Precision} = \frac{\text{True +ve}}{\text{Predicted +ve}}$$

$$\text{Precision} = \frac{\text{True+ve}}{\text{True+ve + False+ve}}$$

$$\text{Accuracy} = \frac{\text{True +ve + True−ve}}{\text{Total}}$$

$$\text{Accuracy} = \frac{\text{True+ve + True −ve}}{\text{True+ve + False−ve + False+ve + True −ve}}$$

$$F_1 \text{ Score} = 2^* \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}}$$

# Confusion Matrix

| Spam filtering | | Predicted | | Total | |
|---|---|---|---|---|---|
| | | Positive | Negative | | |
| Actual | Positive | 952 | 526 | 1478 | Recall(Sensitivity) |
| | Negative | 167 | 3025 | 3192 | Specificity |
| Total | | 1119 | 3551 | 4670 | |
| | | Precision | | | Accuracy, F1 Score |

$$Recall\ (Sensitivity) = \frac{952}{1478} = 0.644$$

$$Specificity = \frac{3025}{3025 + 167} = \frac{3025}{3192} = 0.948$$

$$Precision = \frac{952}{1119} = 0.851$$

Which measure(s) is/are more important?

$$Accuracy = \frac{952 + 3025}{952 + 3025 + 526 + 167} = \frac{3977}{4670} = 0.852$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * 0.851 * 0.644}{0.851 + 0.644} = \frac{1.096}{1.495} = 0.733$$

# Confusion Matrix

| Court System – Death Sentence | | Verdict | | |
|---|---|---|---|---|
| | | Guilty | Not Guilty | |
| Actual | Guilty | True +ve | False –ve | Recall/Sensitivity/True Positive Rate (Minimize False –ve) |
| | Not Guilty | False +ve | True –ve | Specificity/True Negative Rate (Minimize False +ve) |
| | | Precision | | Accuracy, $F_1$ score |

Which measure(s) is/are more important?

# Confusion Matrix

| Breast cancer detection | | Predicted | | Total |
|---|---|---|---|---|
| | | Positive | Negative | |
| Actual | Positive | 852 | 126 | 978 |
| | Negative | 67 | 1025 | 1092 |
| Total | | 919 | 1151 | 2070 |

$$Recall\ (Sensitivity) = \frac{852}{978} = 0.871$$

$$Precision = \frac{852}{919} = 0.927$$

$$Accuracy = \frac{852 + 1025}{852 + 1025 + 126 + 67} = \frac{1877}{2070} = 0.907$$

$$Specificity = \frac{1025}{1025 + 67} = \frac{1025}{1092} = 0.939$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * 0.871 * 0.927}{0.871 + 0.927} = \frac{1.615}{1.798} = 0.898$$

Which measure(s) is/are more important?

# Confusion Matrix

| Anti Virus Detection | | Detection | | |
|---|---|---|---|---|
| | | Virus | No Virus | |
| Actual | Virus | True +ve | False –ve | Recall/Sensitivity/True Positive Rate (Minimize False –ve) |
| | No Virus | False +ve | True –ve | Specificity/True Negative Rate (Minimize False +ve) |
| | | Precision | | Accuracy, $F_1$ score |

Which measure(s) is/are more important?

# Confusion Matrix

| Organ Matching from Donors | | Predicted | | |
|---|---|---|---|---|
| | | Match | No Match | |
| Actual | Match | True +ve | False –ve | Recall/Sensitivity/True Positive Rate (Minimize False –ve) |
| | No Match | False +ve | True –ve | Specificity/True Negative Rate (Minimize False +ve) |
| | | Precision | | Accuracy, $F_1$ score |

Which measure(s) is/are more important?

# Confusion Matrix

| Credit Card Fraud Detection | | Detection | | |
|---|---|---|---|---|
| | | Fraud | No Fraud | |
| Actual | Fraud | True +ve | False –ve | Recall/Sensitivity/True Positive Rate (Minimize False –ve) |
| | No Fraud | False +ve | True –ve | Specificity/True Negative Rate (Minimize False +ve) |
| | | Precision | | Accuracy, $F_1$ score |

Which measure(s) is/are more important?

# Confusion Matrix

| Image Text Classification | | Predicted Word | | |
|---|---|---|---|---|
| | | CAT | DOG | |
| Actual Word | CAT | True +ve | False –ve | Recall/Sensitivity/True Positive Rate (Minimize False –ve) |
| | DOG | False +ve | True –ve | Specificity/True Negative Rate (Minimize False +ve) |
| | | Precision | | Accuracy, $F_1$ score |

Which measure(s) is/are more important?

Analyzing attributes

# PROBABILITY DISTRIBUTIONS

# Random Variable

- A variable that can take multiple values with different probabilities.
- The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.

| **Points scored per game** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|
| Frequency, f | 1 | 4 | 6 | 12 | 5 | 1 | 1 |

Total number of games – N= 30

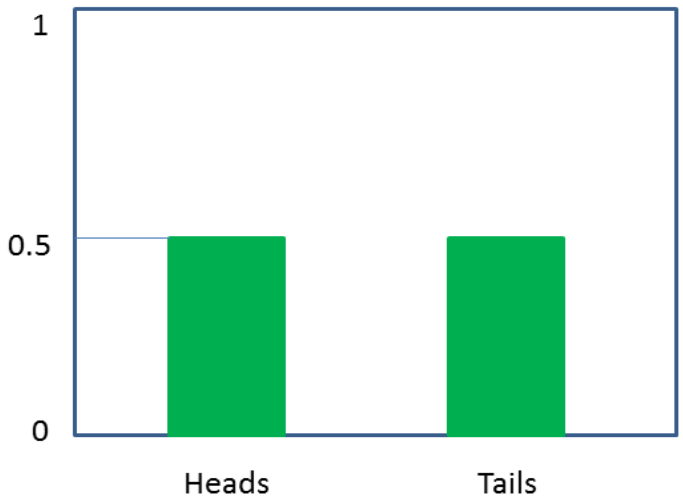
| **Points scored per game** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|
| Probability<br>*Recall the Frequentist (empirical) approach of assigning probabilities* | $\frac{1}{30}$ | $\frac{4}{30}$ | $\frac{6}{30}$ | $\frac{12}{30}$ | $\frac{5}{30}$ | $\frac{1}{30}$ | $\frac{1}{30}$ |

Leads to Descriptive Stats

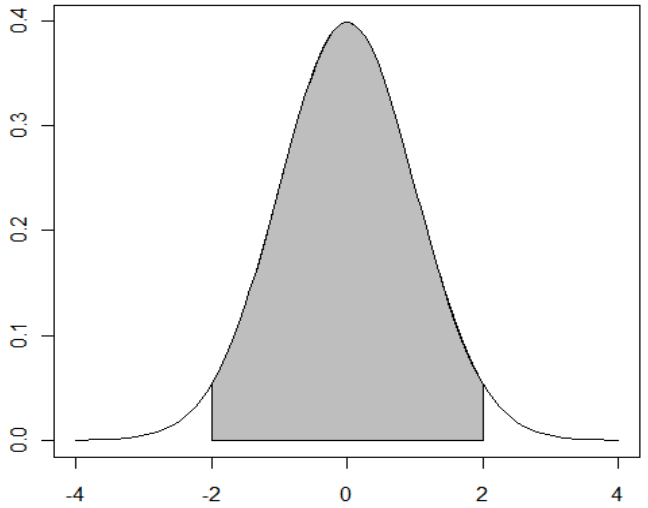
Leads to Inferential Stats

CSE 7315c

# Discrete and Continuous




Countable



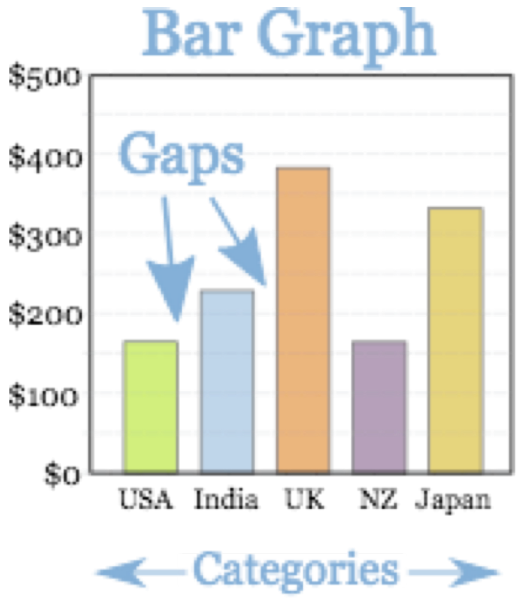

Measurable

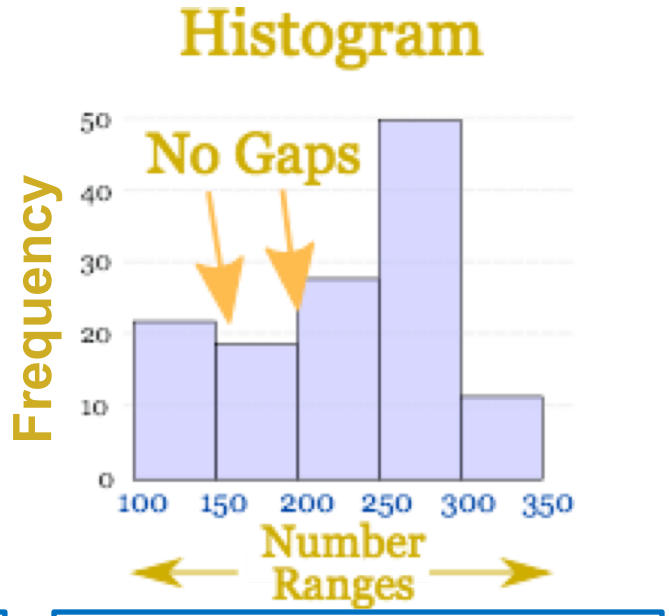
# Can any function be a probability distribution?

| Discrete Distributions | Continuous Distributions |
|---|---|
| Probability that X can take a specific value $x$ is $P(X = x) = p(x)$. | Probability that X is between two points $a$ and $b$ is $P(a \leq X \leq b) = \int_a^b f(x)dx$. |
| It is non-negative for all real $x$. | It is non-negative for all real $x$. |
| The sum of $p(x)$ over all possible values of $x$ is 1, i.e., $\sum p(x) = 1$. | $\int_{-\infty}^{\infty} f(x)dx = 1$ |
| **Probability Mass Function** | **Probability Density Function** |

# Histogram

A series of **contiguous rectangles** that represent the frequency of data in given class intervals.




Gaps have no meaning in Bar Graph

Gaps have significance in Histogram

How many class intervals?

- Rule of thumb: 5-15 (not too many and not too few)
- Freedman-Diaconis rule:

$$No.of\ bins = \frac{(max - min)}{2 * IQR * n^{\frac{-1}{3}}},$$

$$where\ the\ denominator\ is\ the\ bin - width$$

# Histogram - Excel

## Annual traffic data for 30 busiest airports in the world – 2013 and 2011

Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2011-final and http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2013-final

Last accessed: February 04, 2016

| | | | | |
|---|---|---|---|---|
| 1 | Passenger Traffic 2011 FINAL (Annual) | | | |
| 2 | Last Update: 8 July 2013 | | | |
| 3 | Passenger Traffic | | | |
| 4 | Total passengers enplaned and deplaned, passengers in transit counted once | | | |
| 5 | Rank | City (Airport) | Total Passengers | % Change |
| 6 | 1 | ATLANTA GA, US (ATL) | 92389023 | 3.5 |
| 7 | 2 | BEIJING, CN (PEK) | 78675058 | 6.4 |
| 8 | 3 | LONDON, GB (LHR) | 69433565 | 5.4 |
| 9 | 4 | CHICAGO IL, US (ORD) | 66701241 | -0.1 |
| 10 | 5 | TOKYO, JP (HND) | 62584826 | -2.5 |
| 11 | 6 | LOS ANGELES CA, US (LAX) | 61862052 | 4.7 |
| 12 | 7 | PARIS, FR (CDG) | 60970551 | 4.8 |
| 13 | 8 | DALLAS/FORT WORTH TX, US (DFW) | 57832495 | 1.6 |
| 14 | 9 | FRANKFURT, DE (FRA) | 56436255 | 6.5 |
| 15 | 10 | HONG KONG, HK (HKG) | 53328613 | 5.9 |
| 16 | 11 | DENVER CO, US (DEN) | 52849132 | 1.7 |
| 17 | 12 | JAKARTA, ID (CGK) | 51533187 | 16.2 |
| 18 | 13 | DUBAI, AE (DXB) | 50977960 | 8 |
| 19 | 14 | AMSTERDAM, NL (AMS) | 49755252 | 10 |
| 20 | 15 | MADRID, ES (MAD) | 49653055 | -0.4 |
| 21 | 16 | BANGKOK, TH (BKK) | 47910904 | 12 |
| 22 | 17 | NEW YORK NY, US (JFK) | 47644060 | 2.4 |
| 23 | 18 | SINGAPORE, SG (SIN) | 46543845 | 10.7 |
| 24 | 19 | GUANGZHOU, CN (CAN) | 45040340 | 9.9 |
| 25 | 20 | SHANGHAI, CN (PVG) | 41447730 | 2.1 |
| 26 | 21 | SAN FRANCISCO CA, US (SFO) | 40927786 | 4.3 |
| 27 | 22 | PHOENIX AZ, US (PHX) | 40591948 | 5.3 |
| 28 | 23 | LAS VEGAS NV, US (LAS) | 40560285 | 2 |
| 29 | 24 | HOUSTON TX, US (IAH) | 40128953 | -0.9 |
| 30 | 25 | CHARLOTTE NC, US (CLT) | 39043708 | 2.1 |
| 31 | 26 | MIAMI FL, US (MIA) | 38314389 | 7.3 |
| 32 | 27 | MUNICH, DE (MUC) | 37763701 | 8.8 |
| 33 | 28 | KUALA LUMPUR, MY (KUL) | 37704510 | 10.6 |
| 34 | 29 | ROME, IT (FCO) | 37651222 | 3.9 |
| 35 | 30 | ISTANBUL, TR (IST) | 37406025 | 16.3 |

Passenger Traffic 2013 FINAL (Annual)

Last Update: 22 December 2014

Passenger Traffic

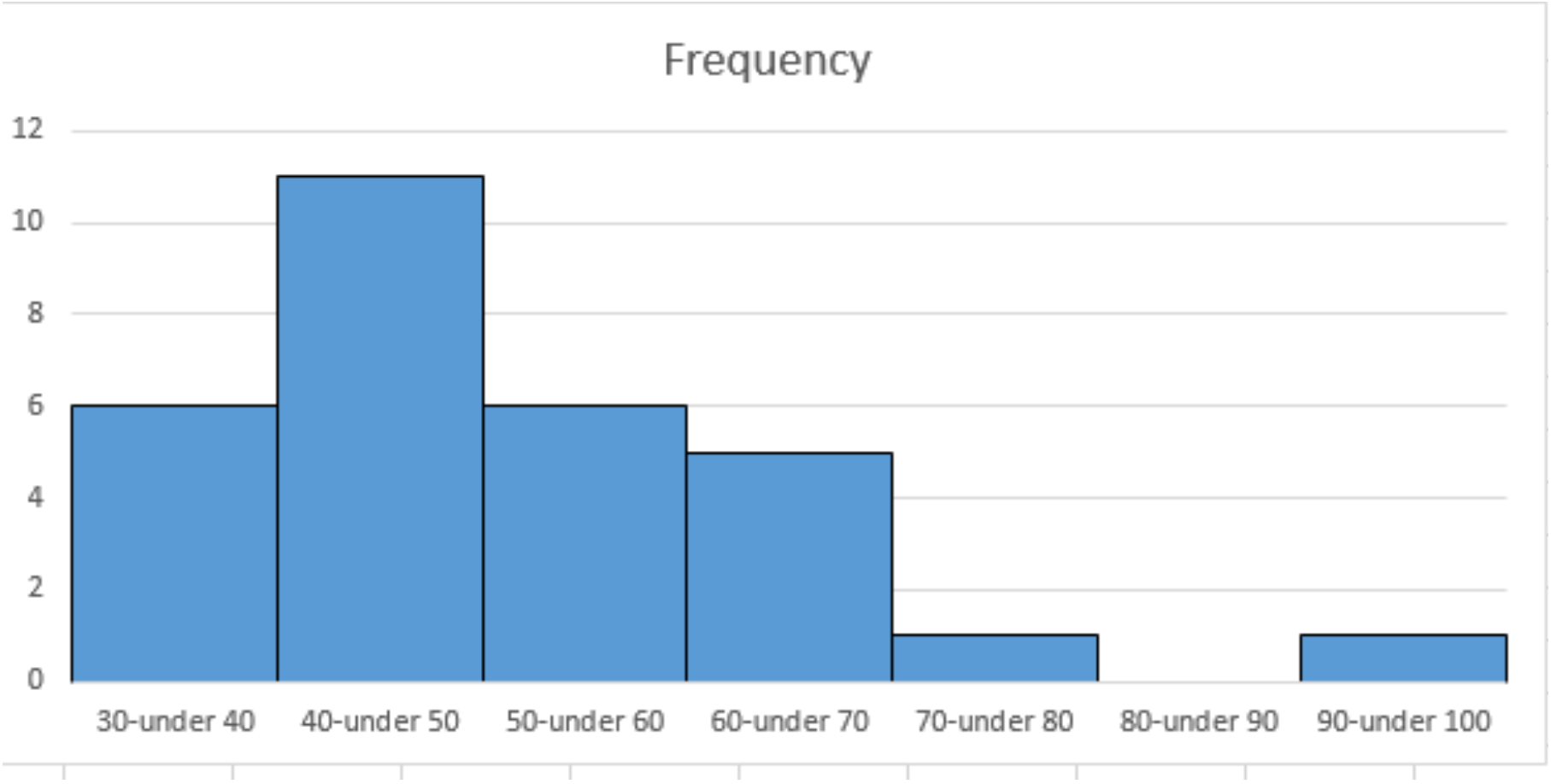
Total passengers enplaned and deplaned, passengers in transit counted once

| Rank | City (Airport) | Passengers 2013 | Passengers 2012 | % Change |
|---|---|---|---|---|
| 1 | ATLANTA GA, US (ATL) | 9,44,31,224 | 9,55,13,828 | -1.1 |
| 2 | BEIJING, CN (PEK) | 8,37,12,355 | 8,19,29,359 | 2.2 |
| 3 | LONDON, GB (LHR) | 7,23,68,061 | 7,00,38,804 | 3.3 |
| 4 | TOKYO, JP (HND) | 6,89,06,509 | 6,67,95,178 | 3.2 |
| 5 | CHICAGO IL, US (ORD) | 6,67,77,161 | 6,66,29,600 | 0.2 |
| 6 | LOS ANGELES CA, US (LAX) | 6,66,67,619 | 6,36,88,121 | 4.7 |
| 7 | DUBAI, AE (DXB) | 6,64,31,533 | 5,76,84,550 | 15.2 |
| 8 | PARIS, FR (CDG) | 6,20,52,917 | 6,16,11,934 | 0.7 |
| 9 | DALLAS/FORT WORTH TX, US (DFW) | 6,04,70,507 | 5,86,20,160 | 3.2 |
| 10 | JAKARTA, ID (CGK) | 6,01,37,347 | 5,77,72,864 | 4.1 |
| 11 | HONG KONG, HK (HKG) | 5,95,88,081 | 5,60,61,595 | 6.3 |
| 12 | FRANKFURT, DE (FRA) | 5,80,36,948 | 5,75,20,001 | 0.9 |
| 13 | SINGAPORE, SG (SIN) | 5,37,26,087 | 5,11,81,804 | 5 |
| 14 | AMSTERDAM, NL (AMS) | 5,25,69,200 | 5,10,35,590 | 3 |
| 15 | DENVER CO, US (DEN) | 5,25,56,359 | 5,31,56,278 | -1.1 |
| 16 | GUANGZHOU, CN (CAN) | 5,24,50,262 | 4,83,09,410 | 8.6 |
| 17 | BANGKOK, TH (BKK) | 5,13,63,451 | 5,30,02,328 | -3.1 |
| 18 | ISTANBUL, TR (IST) | 5,13,04,654 | 4,51,23,758 | 13.7 |
| 19 | NEW YORK NY, US (JFK) | 5,04,23,765 | 4,92,91,765 | 2.3 |
| 20 | KUALA LUMPUR, MY (KUL) | 4,74,98,127 | 3,98,87,866 | 19.1 |
| 21 | SHANGHAI, CN (PVG) | 4,71,89,849 | 4,48,80,164 | 5.1 |
| 22 | SAN FRANCISCO CA, US (SFO) | 4,49,45,760 | 4,43,99,885 | 1.2 |
| 23 | CHARLOTTE NC, US (CLT) | 4,34,57,471 | 4,12,28,372 | 5.4 |
| 24 | INCHEON, KR (ICN) | 4,16,79,758 | 3,91,54,375 | 6.4 |
| 25 | LAS VEGAS NV, US (LAS) | 4,09,33,037 | 4,07,99,830 | 0.3 |
| 26 | MIAMI FL, US (MIA) | 4,05,62,948 | 3,94,67,444 | 2.8 |
| 27 | PHOENIX AZ, US (PHX) | 4,03,41,614 | 4,04,48,932 | -0.3 |
| 28 | HOUSTON TX, US (IAH) | 3,97,99,414 | 3,98,91,444 | -0.2 |
| 29 | MADRID, ES (MAD) | 3,97,17,850 | 4,51,76,978 | -12.1 |
| 30 | MUNICH, DE (MUC) | 3,86,72,644 | 3,83,60,604 | 0.8 |

# Histogram

## Annual traffic data for 30 busiest airports in the world – 2011

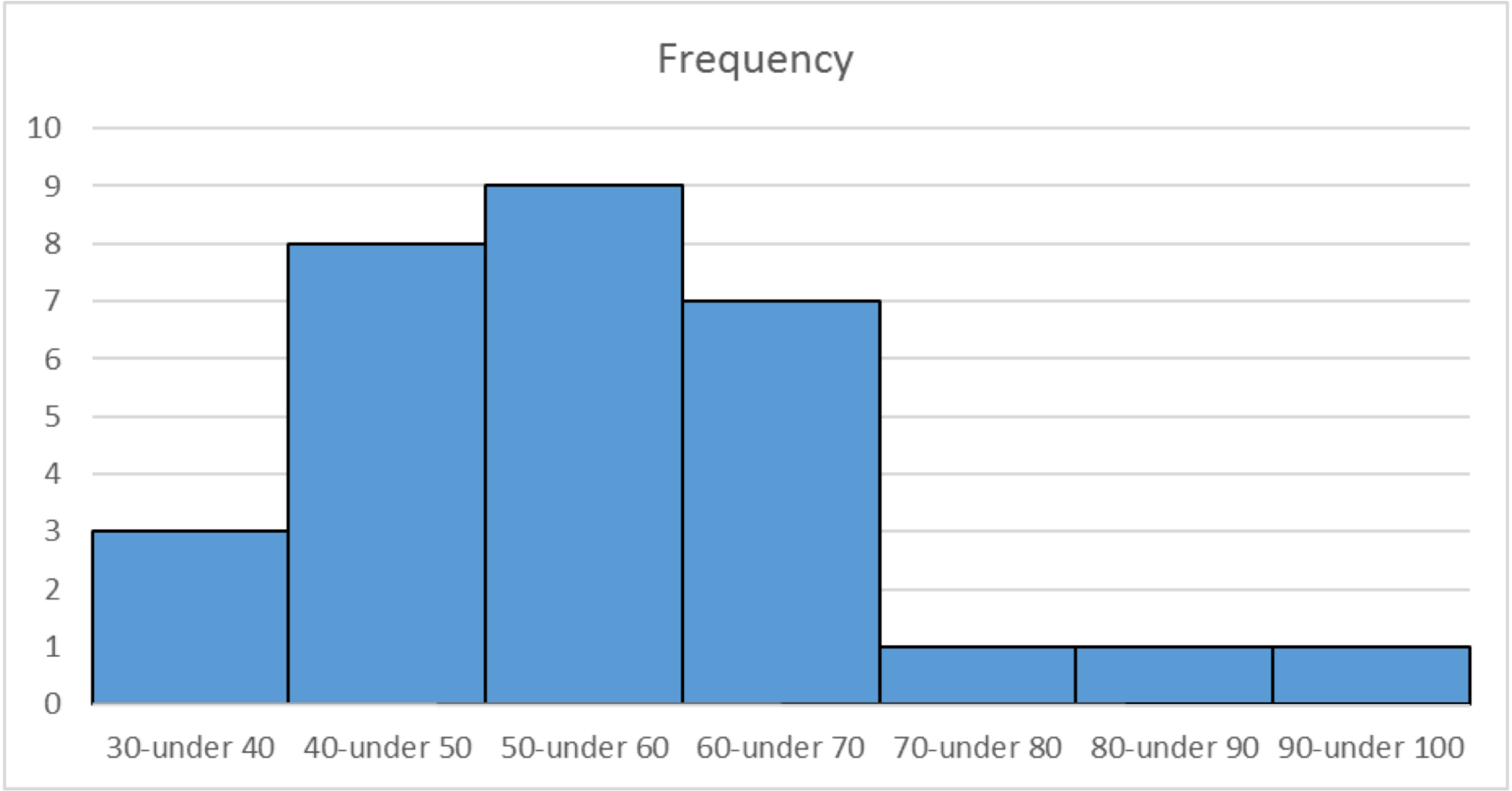
Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2011-final
Last accessed: November 22, 2014



# Histogram

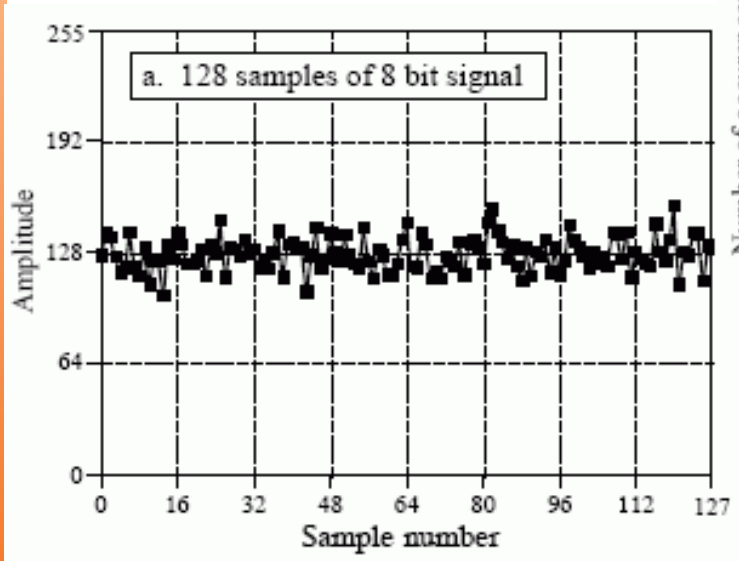
## Annual traffic data for 30 busiest airports in the world – 2013

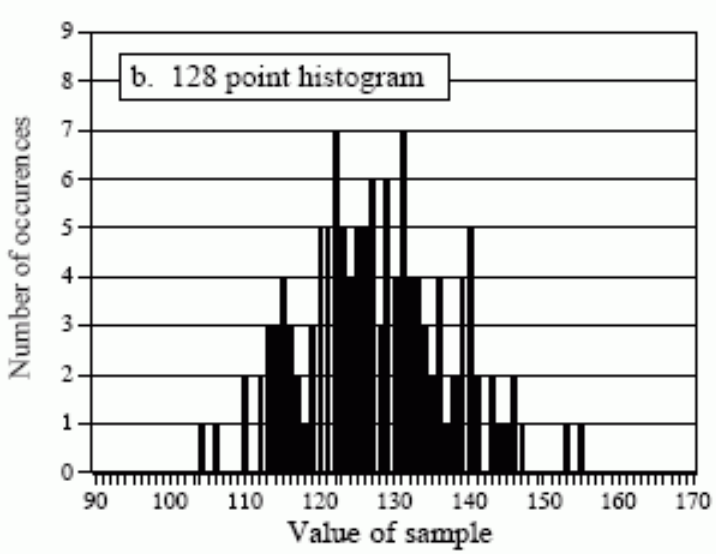
Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2013-final
Last accessed: February 04, 2016

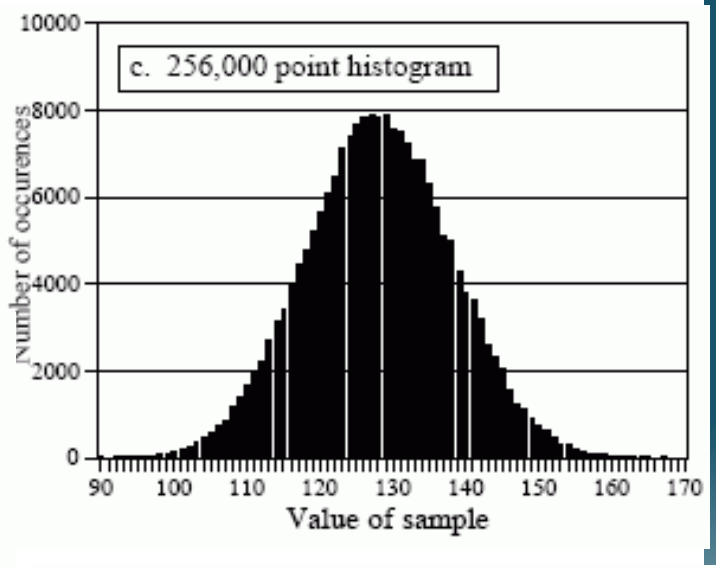


# Histogram, PMF and PDF

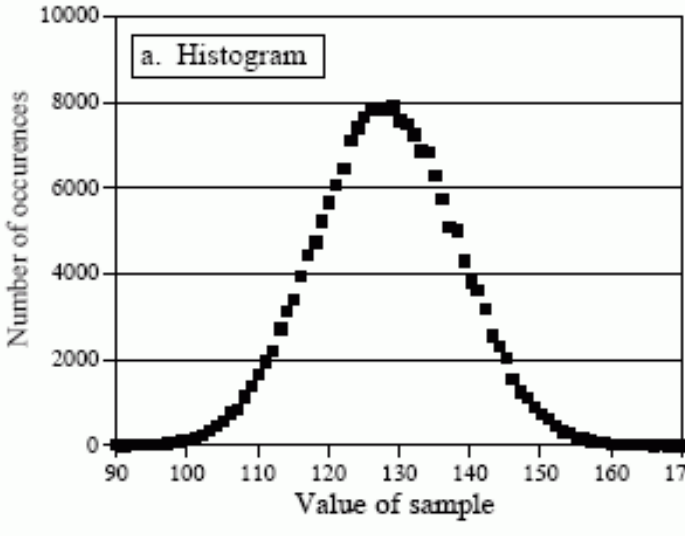
Signal from an 8-bit analog-to-digital converter attached to a computer, e.g., 0-255 mV converted to digital numbers between 0 and 255.



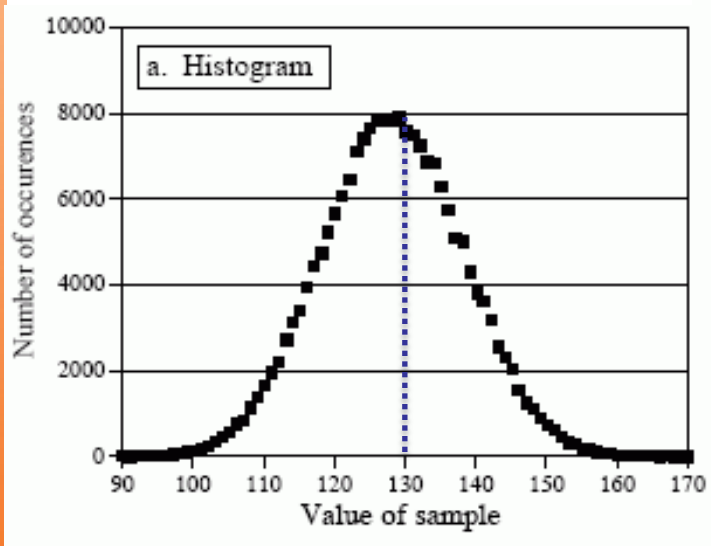

Bin Size for example 100 = 99.5 – 100.5

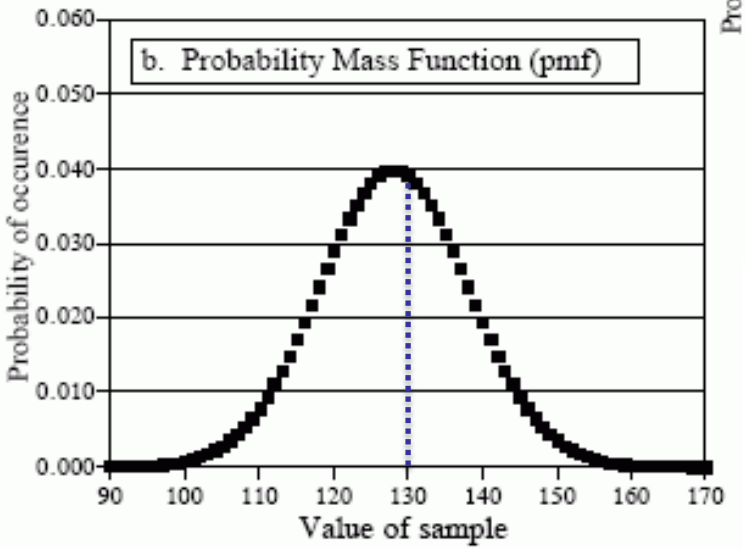


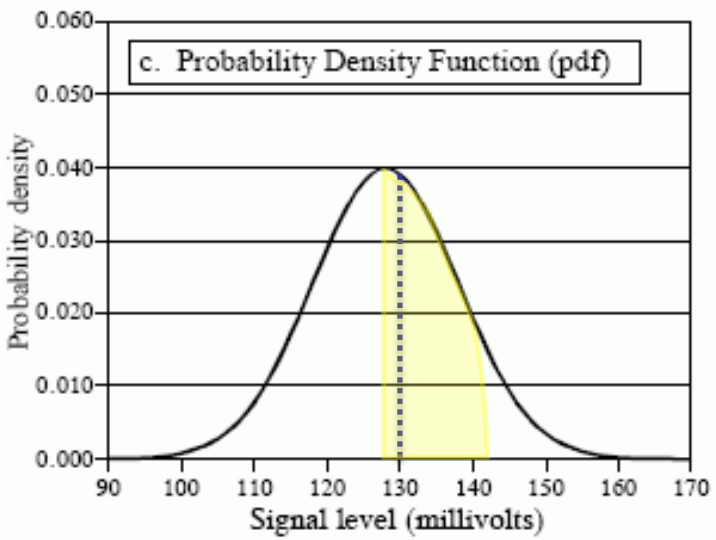

Taking a Analog signal (Continuous Variable ) and Converting it to a Digital signal (Discrete Variable)

# Histogram, PMF and PDF

Signal from an 8-bit analog-to-digital converter attached to a computer, e.g., 0-255 mV converted to digital numbers between 0 and 255.




Have taken 2 Lakhs total samples for calculation instead of 256,000 for easy calculations in PMF and PDF


| Possible Outcome | $ | Cherry | Lemon | Other |
|---|---|---|---|---|
| Probability of Outcome | 0.1 | 0.2 | 0.2 | 0.5 |

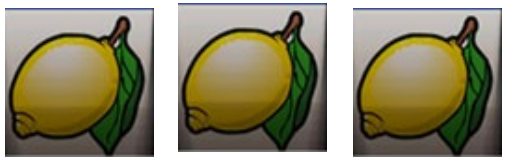
Cost: $1 for each game

Winning combinations:

 = $20

 = $15 (any order)

 = $10

 = $5

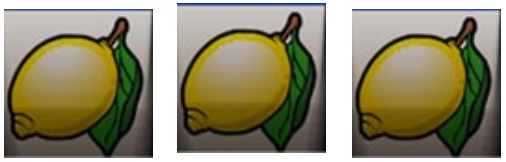
# Probability Distribution of Winnings

Winning combinations:

 = $20.

 = $15 (any order)

 = $10

 = $5

**Cost to play the Game $1**

**Probability of Gain – Remember to subtract the hard earned $1 - All Events are Independent**

**Probability ($20)** = P($ Sign **And** $ Sign **And** $ Sign) = P($ Sign) * P($ Sign) * P($ Sign)

**Probability ($15)** = P($ Sign **And** $ Sign **And** Cherry Sign) in any order = **[**P($ Sign) * P($ Sign) * P(Cherry Sign)**] +** [P($ Sign) * P(Cherry Sign) * P($ Sign)**] + [**P(Cherry Sign) * P($ Sign) * P($ Sign)**]**

**Probability ($10)** = P(Cherry Sign **And** Cherry Sign **And** Cherry Sign) = P(Cherry Sign) * P(Cherry Sign) * P(Cherry Sign)

**Probability ($5)** = P(Lemon Sign **And** Lemon Sign **And** **Lemon** Sign) = P(Lemon Sign) * P(Lemon Sign) * P(Lemon Sign)

**Probability of ($-1)** = 1- P(Of all Winning Combinations)

# Probability Distribution of Winnings

| Combination | None | Lemons | Cherries | Dollars/Cherry | Dollars |
|---|---|---|---|---|---|
| Probability | 0.977 | 0.008 | 0.008 | 0.006 | 0.001 |
| Gain | -$1 | $4 | $9 | $14 | $19 |

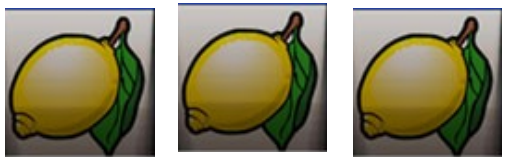
Winning combinations:

**Cost to play the Game $1**

= $20.

= $15 (any order)

= $10



= $5

**Probability of Gain – Remember to subtract the hard earned $1**

Probability ($20) = 0.1*0.1*0.1 = **0.001**

Probability ($15) = (0.1*0.1*0.2)*3 = **0.006**

Probability ($10) = (0.2*0.2*0.2) = **0.008**

Probability ($5) = (0.2*0.2*0.2) = **0.008**

Probability of ($-1) = 1- (0.001+0.006+0.008+0.008) = **0.977**

# Probability Distributions of Winnings and Income

| Combination | None | Lemons | Cherries | Dollars/Cherry | Dollars |
|---|---|---|---|---|---|
| Probability | 0.977 | 0.008 | 0.008 | 0.006 | 0.001 |
| Gain | -$1 | $4 | $9 | $14 | $19 |

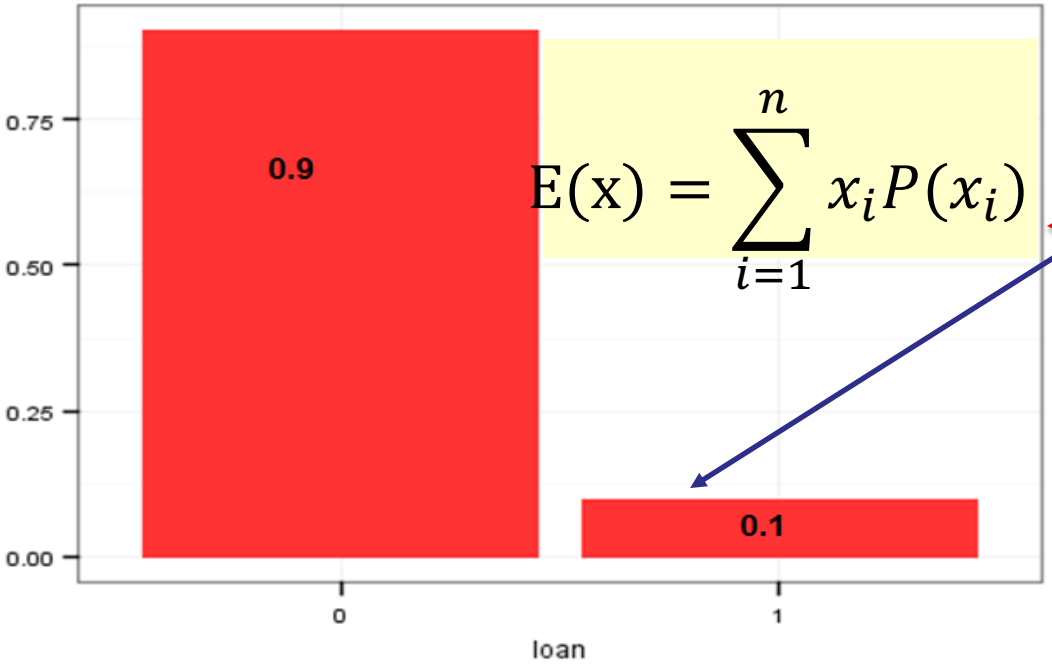
| Salary (BHD) | 100 | 345 | 1000 | 9833 |
|---|---|---|---|---|
| Frequency, f | 10 | 1 | 10 | 2 |
| Probability | 0.43 | 0.04 | 0.43 | 0.09 |

N= 23
P(100) = 10/23 = 0.43
P(345) = 1/23 = 0.04
P(1000) = 10/23 = 0.43
P(9833) = 2/23 = 0.09

Why do you need a probability distribution?

Once a distribution is calculated, it can be used to determine the EXPECTED outcome.

# Expectation: Discrete




$$E(x) = \sum_{i=1}^{n} x_i P(x_i)$$

E(x) = 1*0.1+0.9*0 = 0.1

*Recall anything like this?*

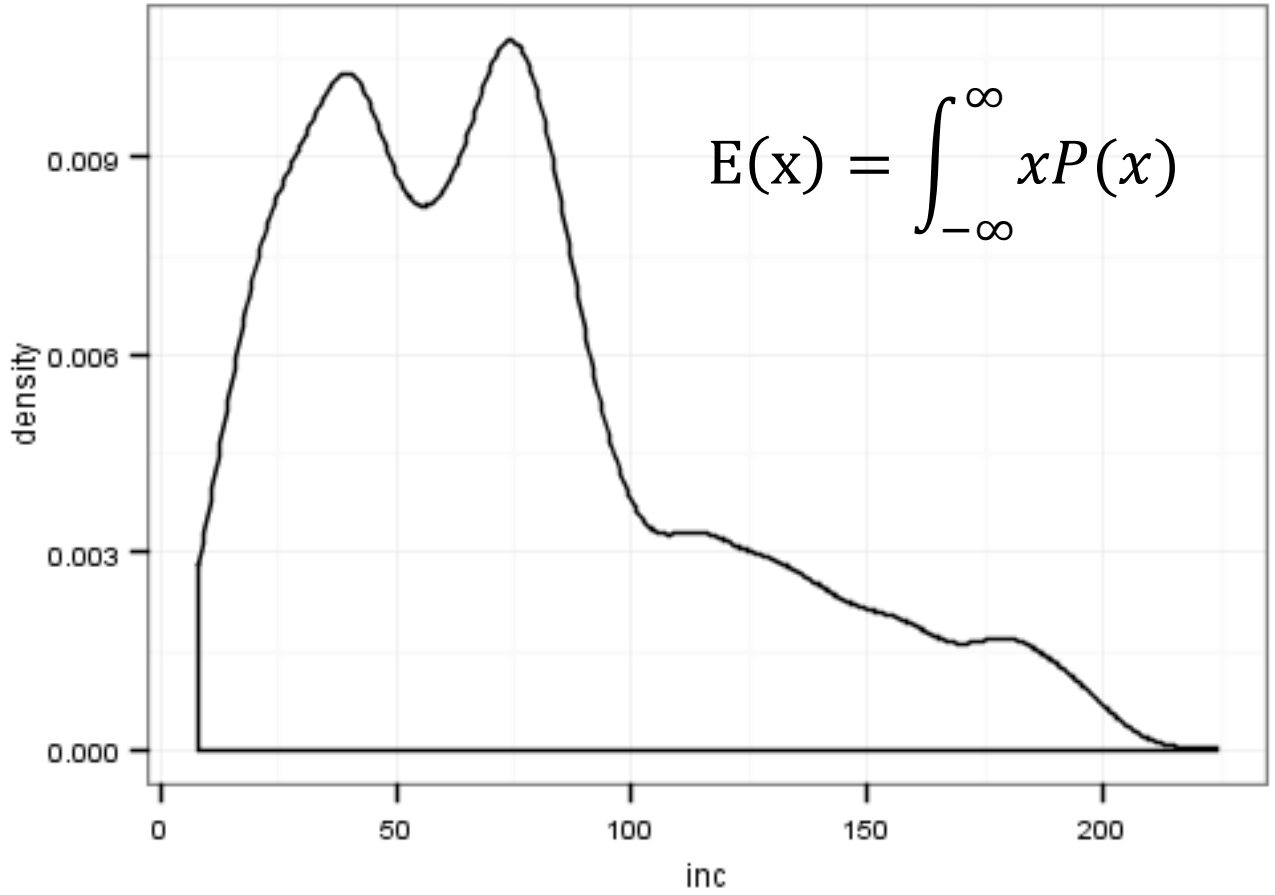
*MEAN*

| Salary (BHD) | 100 | 345 | 1000 | 9833 |
|---|---|---|---|---|
| Frequency, f | 10 | 1 | 10 | 2 |
| Probability | 0.43 | 0.04 | 0.43 | 0.09 |

Mean, $\mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{100X10+345X1+1000X10+9833X2}{10+1+10+2} = 1348$

Expectation, $E(X) = 100 * 0.43 + 345 * 0.04 + 1000 * 0.43 + 9833 * 0.09 = 1348$

# Expectation: Continuous




$$\mathrm{E(x)} = \int_{-\infty}^{\infty} xP(x)$$

# Probability Distribution of Winnings

| Combination | None | Lemons | Cherries | Dollars/Cherry | Dollars |
|---|---|---|---|---|---|
| P(X=x) | 0.977 | 0.008 | 0.008 | 0.006 | 0.001 |
| x | -$1 | $4 | $9 | $14 | $19 |

**EXPECTATION**, $\mathrm{E}(X) = \mu = \Sigma x P(X = x)$

E(X) = 0.977*(-1) + 0.008*4 + 0.008*9 + 0.006 *14 + 0.001* 19

E(X) = -0.77 (**verify**)

This is the amount of $ expected to be "gained" on each pull of the lever.

So, why play?

There is **VARIANCE**.

# Probability Distribution of Winnings

| Combination | None | Lemons | Cherries | Dollars/Cherry | Dollars |
|---|---|---|---|---|---|
| P(X=x) | 0.977 | 0.008 | 0.008 | 0.006 | 0.001 |
| x | -$1 | $4 | $9 | $14 | $19 |

**Variance =** $Var(X) = \frac{\Sigma(x-\mu)^2}{n}$

$$But\ \frac{\Sigma}{n} = \boldsymbol{Expectation}$$

**VARIANCE,** $Var(X) = E(X-\mu)^2 = \Sigma(x-\mu)^2 P(X=x)$

**Standard Deviation,** $\sigma = \sqrt{Var(X)}$

# Simplifying the Formula

$$E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2]$$

$$= E[X^2] - 2\mu E[X] + \mu^2$$ *(we get this as μ is just a number)*

$$= E[X^2] - 2\mu^2 + \mu^2$$

$$= E[X^2] - \mu^2 = E[X^2] - [E(X)]^2$$

# Expectation Properties

E(X+Y) = E(X) + E(Y)  e.g., Playing a game each on 2 slot machines with different probabilities of winning.  This is called **Independent Observation**.

E(aX+b) = aE(X)+E(b) = aE(X) + b  e.g., values x have been changed.  This is called Linear Transformation.

If I have a portfolio of 30% Microsoft, 50% Bank of America and 20% Walmart stocks, the expected return of my portfolio is

E(Portfolio) = 0.3 E(MS) + 0.5 E(BofA) + 0.2 E(Walmart)

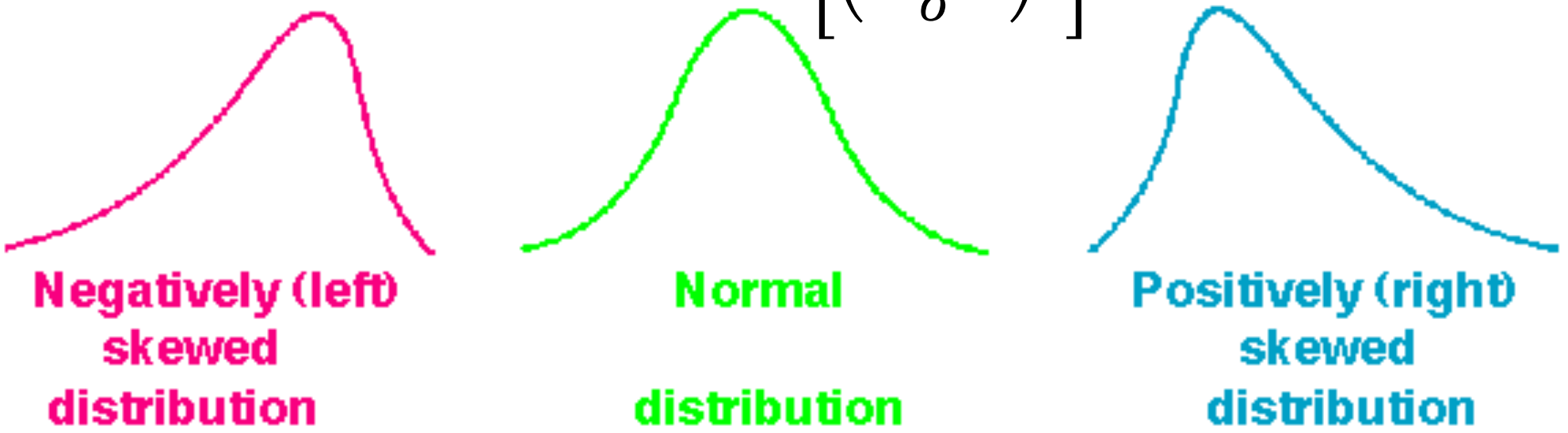
# Variance Properties

- Var(X+a) = Var(X) (Variance does not change when a constant is added)
- Var(X+Y) = Var(X) + Var(Y) for Independent Observations
- Var(X-Y) = Var(X) + Var(Y)
- Var(aX) = $a^2$ Var(X) for **Linear Transformation**

# Understanding the Shape of a PDF - Skewness

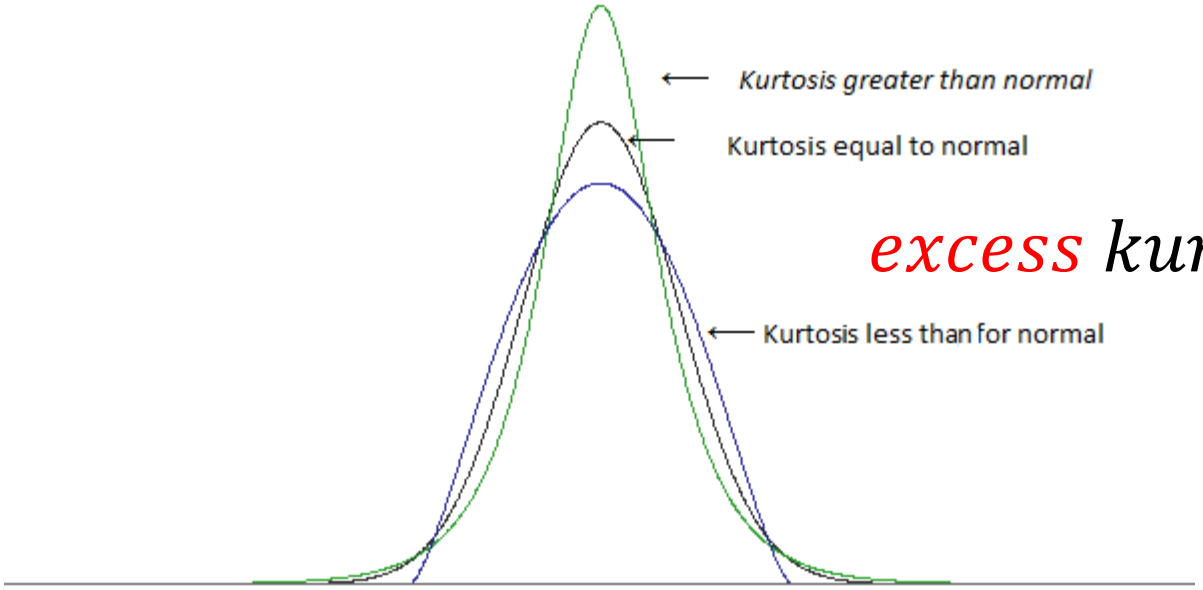
- A measure of symmetry. Negative skew indicates mean is less than median, and positive skew means median is less than mean.

$$skew(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



# Understanding the Shape of a PDF - Kurtosis

A measure of the 'tailed'ness of the data distribution as compared to a normal distribution. Negative kurtosis means a distribution with light tails (fewer extreme deviations from mean (or outliers) than in normal distribution). Positive kurtosis means a distribution with heavy tails (more outliers than in normal distribution).




$$excess\ kurt(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$$

Image Source: http://stats.stackexchange.com/questions/84158/how-is-the-kurtosis-of-a-distribution-related-to-the-geometry-of-the-density-fu
Last accessed: March 31, 2017

# Rules of Thumb – Skewness and Kurtosis

**Skewness**

- Highly skewed: < -1 or > +1
- Moderately skewed: -1 to -0.5 or 0.5 to 1
- Symmetrical: -0.5 to 0.5

**Excess Kurtosis**

- High: < -1 or > +1
- Medium: -1 to -0.5 or 0.5 to 1
- Small: -0.5 to 0.5

# Describing a Distribution – Summary of Moments

BREAK

| Measure | Formula | Description |
|---|---|---|
| Mean ($\mu$) | $E(X)$ | Measures the center of the distribution of X |
| Variance ($\sigma^2$) | $E[(X-\mu)^2]$ | Measures the spread of the distribution of X about the mean |
| Skewness | $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ | Measures asymmetry of the distribution of X |
| Kurtosis (excess) | $E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]-3$ | Measures 'tailed'ness of the distribution of X and useful in outlier identification |

CSE 7315c

# Day 2 : Recall

End of Day 2

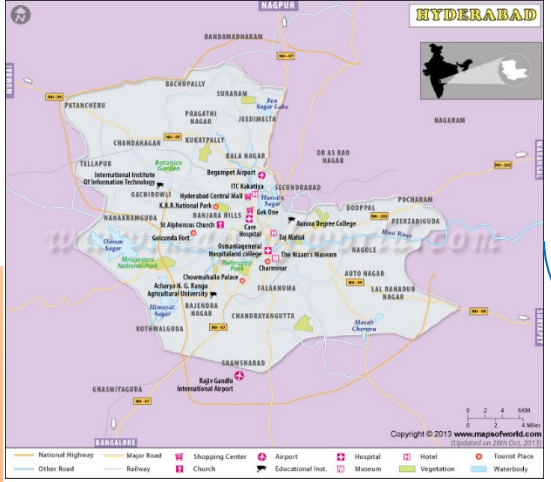
- Interview Questions for Day 1 lesson
- Probability Basics and Types
  - Joint Probability
  - Union Probability
  - Marginal Probability
  - Conditional Probability
- Probability Table and Venn Diagram
- Probability Tree
- Bayes Theorem
- Confusion Matrix
  - Recall (Sensitivity)
  - Precision
  - Accuracy
  - Specificity
  - $F_1$ Score

# Day 2: Recall

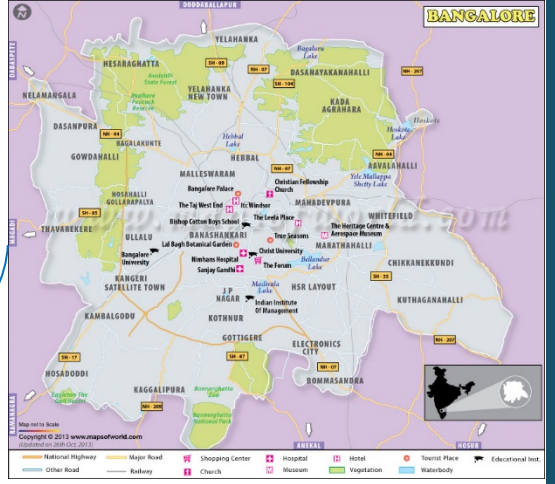
End of Day 2

- Probability Distribution
  - Random Variable (Discrete, Continuous)
  - Histogram
  - Probability Distribution
    - Discrete and Continuous
  - Discrete Probability Distribution – Probability Mass Function (PMF)
  - Continuous Probability Distribution – Probability Density Function (PDF)
  - Expectation and Variance
  - Skewness and Kurtosis

# COMMON DISTRIBUTIONS




Inspire...Educate...Transform.




## HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

## BENGALURU

Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

### Social Media

| | |
|---|---|
| Web: | http://www.insofe.edu.in |
| Facebook: | https://www.facebook.com/insofe |
| Twitter: | https://twitter.com/Insofeedu |
| YouTube: | http://www.youtube.com/InsofeVideos |
| SlideShare: | http://www.slideshare.net/INSOFE |
| LinkedIn: | http://www.linkedin.com/company/international-school-of-engineering |

*This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.*